To appear in the Proc. of the *2020 25th International Conference on Pattern Recognition (ICPR 2020)* Milan, Italy, Jan. 2021.

CausalX: Causal eXplanations and Block Multilinear Factor Analysis

M. Alex O. Vasilescu[1,2]
maov@cs.ucla.edu

Eric Kim[2,1]
ekim@cs.ucla.edu

Xiao S. Zeng[2]
stevennz@ucla.edu

[1]Tensor Vision Technologies, Los Angeles, California
[2]Department of Computer Science,University of California, Los Angeles

***Abstract*—By adhering to the dictum, "No causation without manipulation (treatment, intervention)", cause and effect data analysis represents changes in observed data in terms of changes in the causal factors. When causal factors are not amenable for active manipulation in the real world due to current technological limitations or ethical considerations, a counterfactual approach performs an intervention on the model of data formation. In the case of object representation or activity (temporal object) representation, varying object parts is generally unfeasible whether they be spatial and/or temporal. Multilinear algebra, the algebra of higher order tensors, is a suitable and transparent framework for disentangling the causal factors of data formation. Learning a part-based intrinsic causal factor representations in a multilinear framework requires applying a set of interventions on a part-based multilinear model. We propose a unified multilinear model of wholes and parts. We derive a hierarchical block multilinear factorization, the M-mode Block SVD, that computes a disentangled representation of the causal factors by optimizing simultaneously across the entire object hierarchy. Given computational efficiency considerations, we introduce an incremental bottom-up computational alternative, the Incremental M-mode Block SVD, that employs the lower level abstractions, the part representations, to represent the higher level of abstractions, the parent wholes. This incremental computational approach may also be employed to update the causal model parameters when data becomes available incrementally. The resulting object representation is an interpretable combinatorial choice of intrinsic causal factor representations related to an objects recursive hierarchy of wholes and parts that renders object recognition robust to occlusion and reduces training data requirements.**

***Index Terms*—causality, counterfactuals, explanatory variables, latent representation, factor analysis, multilinear algebra, M-mode SVD, block tensor factorization, hiearchical tensor hierarchical representation, object recognition, image analysis**

I. Introduction: Problem Definition

Developing causal explanations for correct results or for failures from mathematical equations and data is important in developing a trustworthy artificial intelligence, and retaining public trust. Causal explanations are germane to the "right to an explanation" statute [15], [13] *i.e.*, to data driven decisions, such as those that rely on images. Computer graphics and computer vision problems, also known as forward and inverse imaging problems, have been cast as causal inference questions [40], [42] consistent with Donald Rubin's quantitative definition of causality, where "A causes B" means "the effect of A is B", a measurable and experimentally repeatable quantity [14], [17]. Computer graphics may be viewed as addressing analogous questions to forward causal inferencing that addresses the "what if" question, and estimates the change in effects given a delta change in a causal factor. Computer vision may be viewed as addressing analogous questions to inverse causal inferencing that addresses the "why" question [12]. We define inverse causal inference as the estimation of causes given an estimated forward causal model and a set of observations that constrain the solution set.

Natural images are the composite consequence of multiple factors related to scene structure, illumination conditions, and imaging conditions. Multilinear algebra, the algebra of higher-order tensors, offers a potent mathematical framework for analyzing the multifactor structure of image ensembles and for addressing the difficult problem of disentangling the constituent factors, Fig. 2. (Vasilescu and Terzopoulos: TensorFaces 2002 [43], [44], MPCA and MICA 2005 [46], kernel variants [40], Multilinear Projection 2007/2011[47], [41])

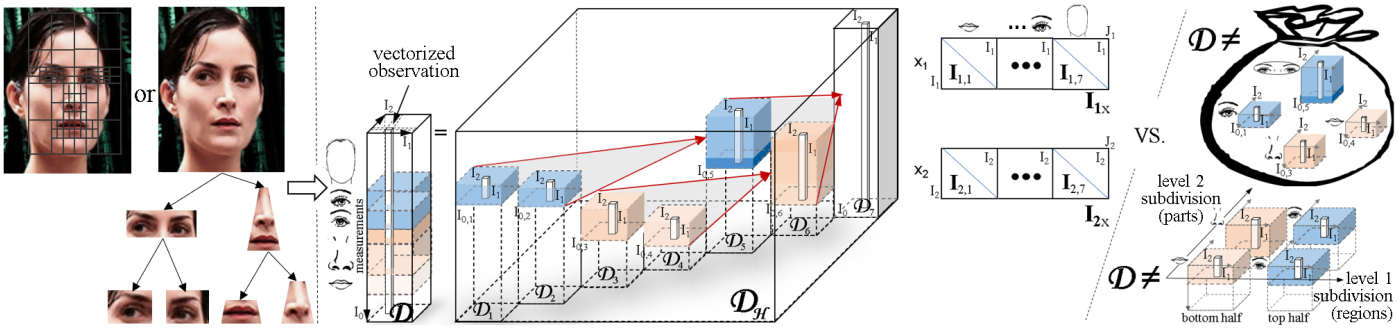

Fig. 1: Data tensor, $\mathcal{D}$, expressed in terms of a hierarchical data tensor, $\mathcal{D}_{\mathcal{H}}$, a mathematical instantiation of a tree data structure where $\mathcal{D} = \mathcal{D}_{\mathcal{H}} \times_1 \mathbf{I}_{1x} \cdots \times_c \mathbf{I}_{cx} \ldots \mathbf{I}_{Cx}$, versus an independent bag of parts/sub-parts, or a data tensor with a reparameterized measurement mode in terms of regions and sub-regions. An object hierarchy may be based on adaptive quad/triangle based subdivision of various depths [38], or a set of perceptual parts of arbitrary shape, size and location. Images of non-articulated objects are best expressed with hierarchical data tensors that have a partially compositional form, where all the parts share the same extrinsic causal factor representations, Fig. 3b. Images of objects with articulated parts are best expressed in terms of hierarchical data tensors that are fully compositional in the causal factors, Fig. 3c. Images of non-articulated objects may also be represented by a fully compositional hierarchical data tensor, as depicted by the TensorTrinity example above.

Scene structure is composed from a set of objects that appear to be formed from a recursive hierarchy of perceptual wholes and parts whose properties, such as shape, reflectance, and color, constitute a hierarchy of intrinsic causal factors of object appearance. Object appearance is the compositional consequence of both an object's intrinsic causal factors, and extrinsic causal factors with the latter related to illumination (i.e. the location and types of light sources), and imaging (i.e. viewing direction, camera lens, rendering style etc.). Intrinsic and extrinsic causal factors confound each other's contributions hindering recognition [42].

"Intrinsic properties are by virtue of the thing itself and nothing else" (David Lewis, 1983 [22]); whereas extrinsic properties are not entirely about that thing, but as a result of the way the thing interacts with the world. Unlike global intrinsic properties, local intrinsic properties are intrinsic to a part of the thing, and it may be said that a local intrinsic property is in an "intrinsic fashion", or "intrinsically" about the thing, rather than "is intrinsic" to the thing [19].

Cause and effect analysis models the mechanisms of data formation, unlike conventional statistical analysis and conventional machine learning that model the distribution of the data [29]. Causal modeling from observational studies are suspect of bias and confounding with some exceptions [8], [34], unlike experimental studies [31], [32] in which a set of active interventions are applied, and their effect on response variables are measured and modeled. The differences between experimental studies, denoted symbolically with Judea Pearl's *do*-operator [29], and observational studies are best exemplified by the following expectation and probability expressions

$$\underbrace{\begin{matrix} E(\mathbf{d}|c) \\ P(\mathbf{d}|c) \end{matrix}}_{\textbf{From Observational Studies:}} \begin{matrix} \neq \\ \neq \end{matrix} \underbrace{\begin{matrix} E(\mathbf{d}|do(c)) \\ P(\mathbf{d}|do(c)), \end{matrix}}_{\textbf{From Experimental Studies:}}$$

Association, Correlation, Prediction — Causation

where $\mathbf{d}$ is a multivariate observation, and c is a hypothesized or actual causal factor. Pearl and Bareinboim [30], [2] have delineated the challenges of generalizing results from experimental studies to observational studies by parameterizing the error based on the possible error inducing sources.

The multilinear (tensor) structural equation approach is a suitable and transparent framework for disentangling the factors of data formation that has been employed in psychometrics [37], [16], [6], [3] econometrics [26],chemometrics [5], [1], signal processing [9], [10], [24], [27] computer vision [44], [11], [49], [50], computer graphics [39], [45], [48], [18], [25], [28], and machine learning [40], [46], [7], [36].

Adhering to the dictum, "No causation without manipulation (treatment, intervention)" [32], [20] each causal factor is varied one at a time while holding the rest fixed, and their effects on the response variables are measured and modeled by a data tensor model. The best evidence comes from randomized comparative studies. However, when causal factors are not amenable for manipulation due to current technological limitations or ethical considerations, a counterfactual approach is required. Rather than performing a manipulation in the real world, a counterfactual approach performs an intervention on the model.

In the case of object representation or activity (temporal object) representation, varying object parts is generally unfeasible whether they be spatial or temporal. Learning a hierarchy of intrinsic causal factor representations requires applying a set of interventions on the structural model, hence it requires a part-based multilinear model, Fig 1.

This paper proposes a unified multilinear model of wholes and parts that defines a data tensor in terms of a *hierarchical data tensor*, $\mathcal{D}_{\mathcal{H}}$, a mathematical instantiation of a tree data structure. Our hierarchical data tensor is a mathematical conceptual device that enables us to derive a hierarchical block multilinear factorization, an M-mode Block SVD, that optimizes simultaneously across the entire object hierarchy and allows for different tree parametrizations for the intrinsic versus the extrinsic causal factors. Given computational considerations, we develop an incremental computational alternative that employs the lower level abstractions, the part representations, to represent the higher level of abstractions, the parent wholes.

Our hierarchical block multilinear factorization, M-mode Block SVD, disentangles the causal structure by computing statistically invariant intrinsic and extrinsic representations. The factorization learns a hierarchy of low-level, mid-level and high-level features. Our hybrid approach mitigates the shortcomings of local features that are sensitive to local deformations and noise, and the shortcomings of global features that are sensitive to occlusions. The resulting object representation is a combinatorial choice of part representations, that renders object recognition robust to occlusion and reduces large training data requirements. This approach was employed for face verification by computing a set of causal explanations (causalX) [42].

II. Relevant Tensor Algebra

We will use standard textbook notation, denoting scalars by lower case italic letters $(a, b, ...)$, vectors by bold lower case letters $(\mathbf{a}, \mathbf{b}, ...)$, matrices by bold uppercase letters $(\mathbf{A}, \mathbf{B}, ...)$, and higher-order tensors by bold uppercase calligraphic letters $(\mathcal{A}, \mathcal{B}, ...)$. Index upper bounds are denoted by italic uppercase (*i.e.*, $1 \leq i \leq I$). The zero matrix is denoted by $\mathbf{0}$, and the identity matrix is denoted by $\mathbf{I}$. References [21], [33] provide a quick tutorial, but references [40], [46], [41] are an indepth treatment of tensor based factor analysis.

Briefly, the natural generalization of matrices (i.e., linear operators defined over a vector space), tensors define multilinear operators over a *set* of vector spaces. A *"data tensor"* denotes an M-way data array.

Definition 1 (Tensor): Tensors are multilinear mappings over a set of vector spaces, $\mathbb{R}^{I_c}$, $1 \leq c \leq C$, to a range vector space $\mathbb{R}^{I_0}$:

$$\mathcal{A} : \left\{ \mathbb{R}^{I_1} \times \mathbb{R}^{I_2} \times \cdots \times \mathbb{R}^{I_C} \right\} \mapsto \mathbb{R}^{I_0}. \tag{1}$$

The *order* of tensor $\mathcal{A} \in \mathbb{R}^{I_0 \times I_1 \times \cdots \times I_C}$ is $M = C + 1$. An element of $\mathcal{A}$ is denoted as $\mathcal{A}_{i_0 i_1 \ldots i_c \ldots i_C}$ or $a_{i_0 i_1 \ldots i_c \ldots i_C}$, where $1 \leq i_0 \leq I_0$, and $1 \leq i_c \leq I_c$.

The mode-m vectors of an M-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ are the I_m-dimensional vectors obtained from $\mathcal{A}$ by varying index i_m while keeping the other indices fixed. In tensor terminology, column vectors are the mode-1 vectors and row vectors as mode-2 vectors. The mode-m vectors of a tensor are also known as *fibers*. The mode-m vectors are the column vectors of matrix $\mathbf{A}_{[m]}$ that results from *matrixizing* (a.k.a. *flattening*) the tensor $\mathcal{A}$.

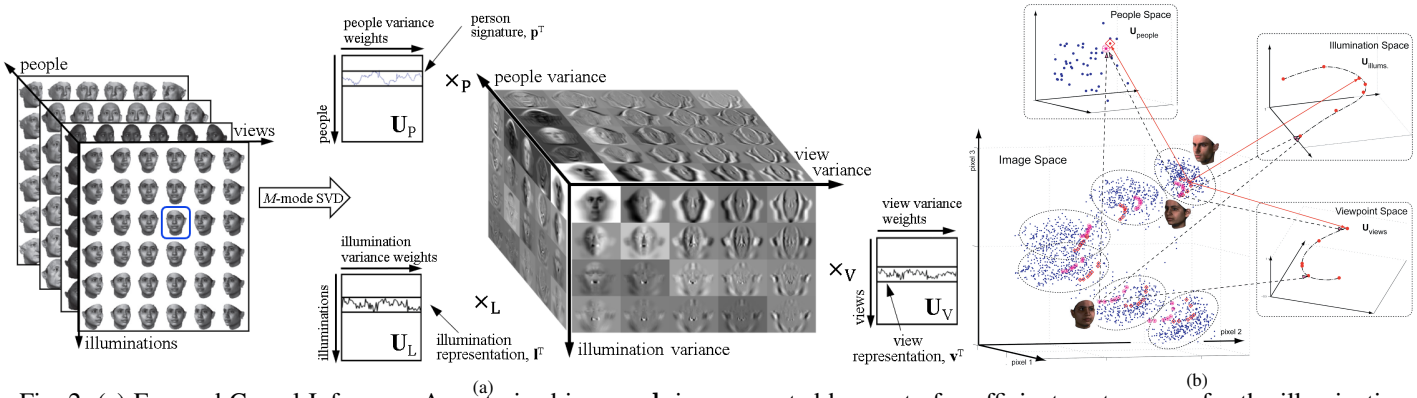

Fig. 2: (a) Forward Causal Inference: A vectorized image, $\mathbf{d}$, is represented by a set of coefficient vectors, one for the illumination, the viewing conditions, and the person $(\mathbf{l}, \mathbf{v}, \mathbf{p})$ and expressed mathematically by $\mathbf{d} = \mathcal{T} \times_{\text{L}} \mathbf{l}^{\text{T}} \times_{\text{V}} \mathbf{v}^{\text{T}} \times_{\text{P}} \mathbf{p}^{\text{T}}$. The TensorFaces basis, $\mathcal{T}$, governs the interaction between the causal factors of data formation [46]. (For display only, the mean was added back.) (b) Inverse Causal Inference: While TensorFaces (Multilinear-PCA or Multilinear-ICA [46]) learns the interaction between the causal factors from training data, it does not prescribe an approach for estimating the causal factors from one or more unlabeled test images that need to be enrolled or recognized. For an unlabeled vectorized test image $\mathbf{d}_{\text{new}}$, the causal factor labels are estimated through a multilinear projection algorithm [47], [41] that is succinctly expressed as the M-mode SVD/CP $\left(\mathcal{T}^{+_{\text{x}}} \times_{\text{x}}^{\text{T}} \mathbf{d}_{\text{new}}\right) \approx \mathbf{r}_{\text{L}} \circ \mathbf{r}_{\text{V}} \circ \mathbf{r}_{\text{P}}$ where $\mathbf{r}_{\text{L}}$, $\mathbf{r}_{\text{V}}$, and $\mathbf{r}_{\text{P}}$, are the estimated latent representations from which the view, illumination and person label may be inferred.

Definition 2 (Mode-m Matrixizing): The mode-m matrixizing of tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_M}$ is defined as the matrix $\mathbf{A}_{[m]} \in \mathbb{R}^{I_m \times (I_1 \ldots I_{m-1} I_{m+1} \ldots I_M)}$. As the parenthetical ordering indicates, the mode-m column vectors are arranged by sweeping all the other mode indices through their ranges, with smaller mode indexes varying more rapidly than larger ones; thus,

$$[\mathbf{A}_{[m]}]_{jk} = a_{i_1 \ldots i_m \ldots i_M}, \quad \text{where} \tag{2}$$

$$j = i_m \quad \text{and} \quad k = 1 + \sum_{\substack{n=0 \\ n \neq m}}^{M} (i_n - 1) \prod_{\substack{l=0 \\ l \neq m}}^{n-1} I_l.$$

A generalization of the product of two matrices is the product of a tensor and a matrix [9].

Definition 3 (Mode-m Product, $\times_m$): The mode-m product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_m \times \cdots \times I_M}$ and a matrix $\mathbf{B} \in \mathbb{R}^{J_m \times I_m}$, denoted by $\mathcal{A} \times_m \mathbf{B}$, is a tensor of dimensionality $\mathbb{R}^{I_1 \times \cdots \times I_{m-1} \times J_m \times I_{m+1} \times \cdots \times I_M}$ whose entries are computed by

$$[\mathcal{A} \times_m \mathbf{B}]_{i_1 \ldots i_{m-1} j_m i_{m+1} \ldots i_M} = \sum_{i_m} a_{i_1 \ldots i_{m-1} i_m i_{m+1} \ldots i_M} b_{j_m i_m},$$

$$\mathcal{C} = \mathcal{A} \times_m \mathbf{B}. \quad \underset{\text{tensorize}}{\overset{\text{matrixize}}{\Longleftrightarrow}} \quad \mathbf{C}_{[m]} = \mathbf{B}\mathbf{A}_{[m]}.$$

The M-mode SVD (aka. the Tucker decomposition) is a "generalization" of the conventional matrix (i.e., 2-mode) SVD which may be written in tensor notation as

$$\mathbf{D} \quad = \mathbf{U}_1 \mathbf{S} \mathbf{U}_2^{\text{T}} \qquad \Leftrightarrow \qquad \mathbf{D} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2. \tag{3}$$

The M-mode SVD orthogonalizes the M spaces and decomposes the tensor as the *mode-m product*, denoted $\times_m$, of M-orthonormal mode matrices, and a core tensor $\mathcal{Z}$

$$\mathcal{D} \quad = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \cdots \times_m \mathbf{U}_m \cdots \times_M \mathbf{U}_M. \tag{4}$$

III. Hierarchical Block Tensor Factorizations of $\mathcal{D}$

Within the tensor mathematical framework, a M-way array or "data-tensor", $\mathcal{D} \in \mathbb{R}^{I_0 \times I_1 \cdots \times I_c \cdots \times I_C}$ contains a collection of vectorized and centered observations,[1] $\mathbf{d}_{i_1 \ldots i_c \ldots i_C} \in \mathbb{R}^{I_0}$ that are the result of C causal factors. The c causal factor $(1 \leq c \leq C)$ takes one of I_c values that are indexed by i_c, $1 \leq i_c \leq I_c$. An observation that is result of the confluence C causal factors is modeled by a multilinear structural equation with multimode latent variables, $\mathbf{r}_{\text{c}}$, that represent the causal factors

$$\mathbf{d}_{i_1, \ldots, i_c, \ldots, i_C} = \mathcal{T} \times_1 \mathbf{r}_1{}^{\text{T}} \cdots \times_c \mathbf{r}_c{}^{\text{T}} \cdots \times_C \mathbf{r}_C{}^{\text{T}} + \boldsymbol{\epsilon}_{i_1, \ldots, i_c, \ldots, i_C}, \tag{5}$$

where $\mathcal{T} = \mathcal{Z} \times_0 \mathbf{U}_0$ is the extended core which modulates the interaction between the latent variables, $\mathbf{r}_{\text{c}}$, that represent the causal factors and $\boldsymbol{\epsilon}_{i_1, \ldots, i_c, \ldots, i_C} \in \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is an additive identically and independently distributed (IID) Gaussian noise, Fig. 2.

A. Hierarchical Data Tensor, $\mathcal{D}_{\mathcal{H}}$

We identify a general base case object and two special cases. A base case object may be composed of (i) two partially overlapping children-parts and parent-whole that has data not contained in any of the children-parts, (ii) a set of non-overlapping parts, or (iii) a set of fully overlapping parts. The tensor representation of an object with fully overlapping parts, Fig. 3(e), resembles the rank-(L, M, N) or a rank-$(L, M, \cdot)$ block tensor decomposition [10].[2]

The data wholes and parts are extracted by employing a filter bank $\{\mathbf{H}_{\text{s}} \in \mathbb{R}^{I_0 \times I_0} \mid \sum_{s=1}^{S} \mathbf{H}_{\text{s}} = \mathbf{I}, \text{ and } 1 \leq s \leq S\}$ where a $1D$ ($2D$ or $3D$) convolutional filter, $\mathbf{h}_{\text{s}}$, is written as a circulant matrix (doubly or triply circulant matrix), $\mathbf{H}_{\text{s}}$, and s is a segment index. The convolution may be written as a matrix-vector multiplication or the mode-m product, $\times_{\text{m}}$, between a circulant matrix, $\mathbf{H}_{\text{s}}$ and a vectorized observation. For example, if an observation is returned by the capture device as a 2-way

[1] Reference [40, Appendix A] evaluates some of the arguments found in highly cited publications in favor of treating an image as a matrix rather than vectorizing it. While technically speaking it is not incorrect to treat an image as a matrix, the evaluation concludes that most arguments do not stand up to analytical scrutiny, and it is preferable to vectorize images.

[2] The block tensor decomposition [10] goal is to find the best fitting K fully overlaping tensor blocks that are all multilinearly decomposable into the same multilinear rank-(R_1, R_2, R_3). This is analogous to finding the best fitting K rank-1 terms (also known as rank-$(1, 1, 1)$) computed by the CP-algorithm.

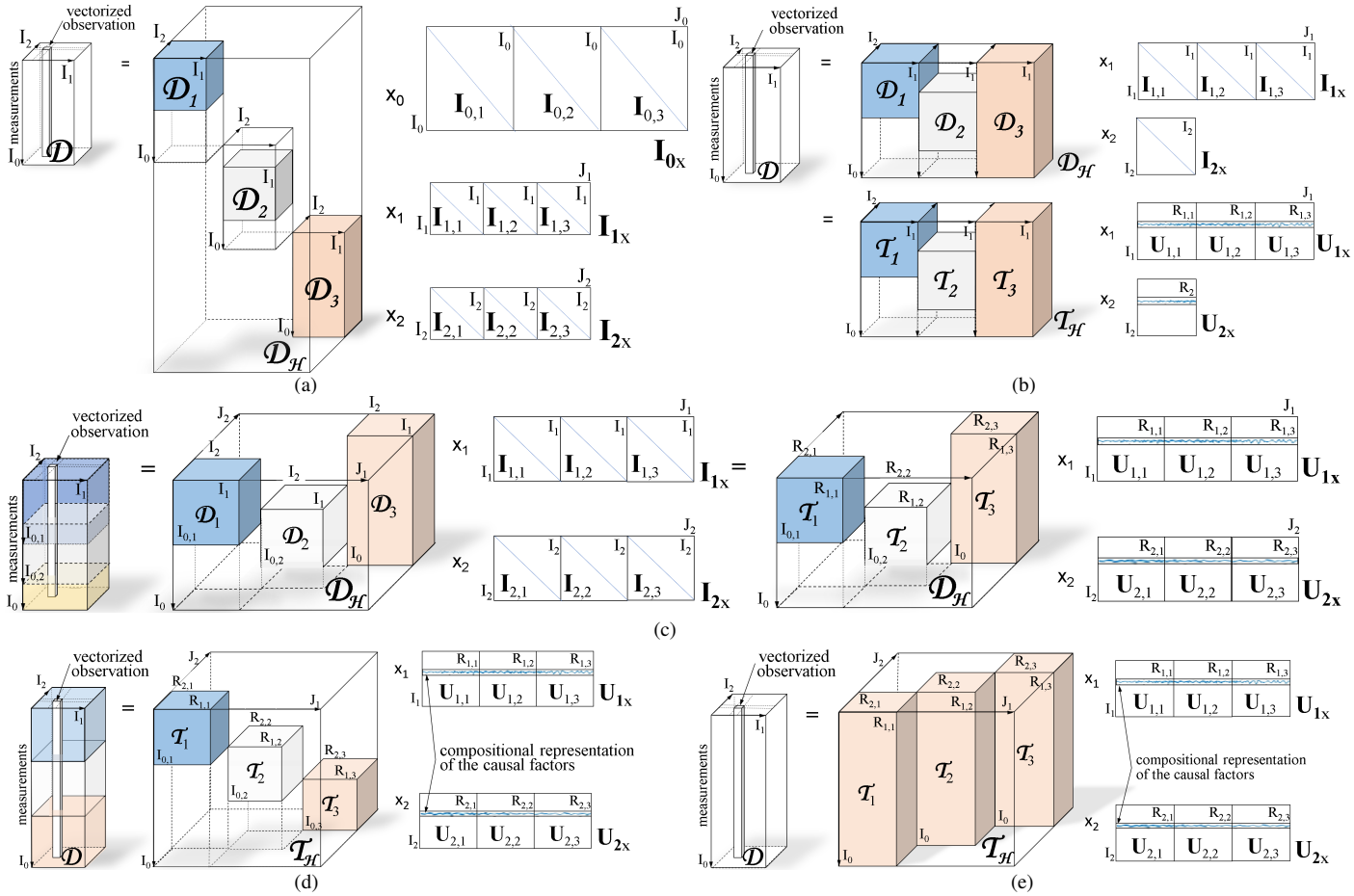

Fig. 3: The data tensor, $\mathcal{D}$, written in terms of a hierarchical data tensor, $\mathcal{D}_{\mathcal{H}}$. (a) When $\mathcal{D}_{\mathcal{H}}$ contains the data tensor segments,$\mathcal{D}_s$, along its super-diagonal then $\mathcal{D}_{\mathcal{H}}$ has a *fully compositional* form, and every mode matrix has a compositional representation. (b) A general base case object written in a *partially compositional* form with a compositional representation for only one mode matrix (causal factor). (c) A general base case object where all the causal factors have a compositional representation. The tensor $\mathcal{D}_{\mathcal{H}}$ is *fully compositional in the causal factors*.(d) A base case object with non-overlapping parts. All the causal factors have a compositional representations. Multilinear factorizations are block independent. (e) Base case object with completely overlapping parts where all the causal factors have compositional representation. Objects with non-overlapping or completely overlapping parts may also be written using a partially compositional form analogously to (b).

multivariate array, $\mathbf{D} \in \mathbb{R}^{I_{\text{xr}} \times I_{\text{xc}}}$, with I_{xr} rows and I_{xc} columns, the convolution is written as

$$\mathbf{D}_{\text{s}} = \mathbf{D} * \mathbf{h}_s(x,y) \underset{\text{matrixize}}{\overset{\text{vectorize}}{\rightleftarrows}} \mathbf{d}_{\text{s}} = \mathbf{H}_{\text{s}}\,\mathbf{d} = \mathbf{d} \times_0 \mathbf{H}_{\text{s}} \tag{6}$$

where the measurement mode is mode 0. In practice, a convolution is efficiently implemented using a DFFT. The segment data tensor, $\mathcal{D}_s = \mathcal{D} \times_0 \mathbf{H}_{\text{s}}$, is the result of multiplying (convolving) every observation, $\mathbf{d}$, with the block circulant matrix (filter), $\mathbf{H}_{\text{s}}$ ($\mathbf{h}_{\text{s}}$). A filter $\mathbf{H}_{\text{s}}$ may be of any type, and have any spatial scope. When a filter matrix is a block identity matrix, $\mathbf{H}_{\text{s}} = \mathbf{I}_{\text{s}}$, the filter matrix multiplication with a vectorized observation has the effect of segmenting a portion of the data. Measurements associated with perceptual parts may not be tightly packed into a block apriori, as in the case of vectorized images, but chunking is achieved by a trivial permutation.

A data tensor is expressed as a recursive hierarchy of wholes and parts by defining and employing a *hierarchical data tensor*, $\mathcal{D}_{\mathcal{H}}$.When a data tensor contains along its super-diagonal the data tensor segments, $\mathcal{D}_{\text{s}}$, then $\mathcal{D}_{\mathcal{H}}$ has a *fully compositional* form, and all the data tensor modes have a compositional representation, Fig. 3(a). The data tensor segments, $\mathcal{D}_s$, may be sparse and represent local parts, or may be full and correspond to a filtered version of a parent-whole, as in the case of a Laplacian pyramid. Mathematically writing $\mathcal{D}$ in terms of $\mathcal{D}_{\mathcal{H}}$ is expressed with

$$\begin{aligned}
\mathcal{D} &= \sum_{s=1}^{S} \mathcal{D} \times_0 \mathbf{H}_{\text{s}} && (7)\\
&= \mathcal{D}_1 \cdots + \mathcal{D}_{\text{s}} \cdots + \mathcal{D}_{\text{S}} && (8)\\
&= \mathcal{D}_{\mathcal{H}} \times_0 \mathbf{I}_{\text{Ox}} \times_1 \mathbf{I}_{\text{1x}} \cdots \times_{\text{c}} \mathbf{I}_{\text{cx}} \cdots \times_{\text{C}} \mathbf{I}_{\text{Cx}}, && (9)
\end{aligned}$$

where $\mathbf{I}_{\text{cx}} = [\mathbf{I}_{\text{c},1} \ldots \mathbf{I}_{\text{c,s}} \ldots \mathbf{I}_{\text{c,S}}] \in \mathbb{R}^{I_{\text{c}} \times SI_{\text{c}}}$ is a concatenation of S identity matrices, one for each data segment. In practice, the measurement mode will not be written in compositional form, ie. the multipication with $\mathbf{I}_{0\text{x}}$ would have been carried out. The resulting $\mathcal{D}_{\mathcal{H}}$ is *fully compositional in the causal factors*, where every causal factor has a compositional representation rather than every mode Fig. 3(c). Articulated-objects have parts with their own extrinsic causal factors and benefit from a compositional representation of every causal factor. A non-articulated object where the wholes, and parts share the same extrinsic causal factor representations (same illumination/viewing conditions) benefit from being written in terms of *a partially compositional*

data tensor, where a single factor has a compositional form, the intrinsic object representation, Fig. 3(b). Thus, the $\mathcal{D}_{\mathcal{H}}$ is multiplied through by all the $\mathbf{I}_{\mathrm{cx}}$ except one. Each multiplied $\mathbf{I}_{\mathrm{cx}}$ is replaced by a single place holder identity matrix in the model.

The three different ways of rewriting $\mathcal{D}$ in terms of a hierarchy of wholes and parts, eq. 7-9, results in three mathematically equivalent representations [3] based on factorizing $\mathcal{D}$, $\mathcal{D}_{\mathrm{s}}$ and $\mathcal{D}_{\mathcal{H}}$:

$$\mathcal{D} = \sum_{s=1}^{S} \underbrace{(\mathcal{Z} \times_0 \mathbf{U}_0 \times_1 \mathbf{U}_1 \cdots \times_{\mathrm{c}} \mathbf{U}_{\mathrm{c}} \cdots \times_{\mathrm{C}} \mathbf{U}_{\mathrm{C}})}_{\mathcal{D}} \times_0 \mathbf{H}_{\mathrm{s}} \quad (10)$$

$$= \sum_{s=1}^{S} \underbrace{(\mathcal{Z}_{\mathrm{s}} \times_0 \mathbf{U}_{0,\mathrm{s}} \times_1 \mathbf{U}_{1,\mathrm{s}} \cdots \times_{\mathrm{c}} \mathbf{U}_{\mathrm{c,s}} \cdots \times_{\mathrm{C}} \mathbf{U}_{\mathrm{C,s}})}_{\mathcal{D}_{\mathrm{s}}} \quad (11)$$

$$= \underbrace{(\mathcal{Z}_{\mathcal{H}} \times_0 \mathbf{U}_{0\mathcal{H}} \times_1 \mathbf{U}_{1\mathcal{H}} \cdots \times_{\mathrm{c}} \mathbf{U}_{\mathrm{c}\mathcal{H}} \cdots \times_{\mathrm{C}} \mathbf{U}_{\mathrm{C}\mathcal{H}})}_{\mathcal{D}_{\mathcal{H}}} \times_0 \mathbf{I}_{0\mathrm{x}} \times_1 \mathbf{I}_{1\mathrm{x}} \cdots \times_{\mathrm{c}} \mathbf{I}_{\mathrm{cx}} \cdots \times_{\mathrm{C}} \mathbf{I}_{\mathrm{Cx}} \quad (12)$$

Despite the prior mathematical equivalence, equations 7,10, and equations 8,12 are not flexible enough to explicitly indicate if the parts are organized in a partially compositional form, or a fully compositional form.

The expression of $\mathcal{D}$ in terms of a hierarchical data tensor is a mathematical conceptual device, that enables a unified mathematical model of wholes and parts that can be expressed completely as a mode-m product (tensor-matrix multiplication) and whose factorization can be optimized in a principled manner.

Dimensionality reduction of the compositional representation is performed by optimizing

$$e = \|\mathcal{D} - (\bar{\mathcal{Z}}_{\mathcal{H}} \times_0 \bar{\mathbf{U}}_{0\mathcal{H}} \times_1 \bar{\mathbf{U}}_{1\mathcal{H}} \ldots \times_{\mathrm{C}} \bar{\mathbf{U}}_{\mathrm{C}\mathcal{H}}) \times_0 \mathbf{I}_{0\mathrm{x}} \ldots \times_{\mathrm{C}} \mathbf{I}_{\mathrm{Cx}}\|^2 + \sum_{c=0}^{C} \lambda_{\mathrm{c}} \|\bar{\mathbf{U}}_{\mathrm{c}\mathcal{H}}^{\mathrm{T}} \bar{\mathbf{U}}_{\mathrm{c}\mathcal{H}} - \mathbf{I}\|^2 \quad (13)$$

where $\bar{\mathbf{U}}_{\mathrm{c}\mathcal{H}}$ is the composite representation of the c^{th} mode, and $\mathcal{Z}_{\mathcal{H}}$ governs the interaction between causal factors.Our optimization may be initialized by setting $\mathcal{Z}_{\mathcal{H}}$ and $\mathbf{U}_{\mathrm{c}\mathcal{H}}$ to the M-mode SVD of $\mathcal{D}_{\mathcal{H}}$,[4] [5] and performing dimensionality reduction through truncation, where $\bar{\mathbf{U}}_{\mathrm{c}\mathcal{H}} \in \mathbb{R}^{SI_{\mathrm{c}} \times \bar{J}_{\mathrm{c}}}$, $\bar{\mathcal{Z}}_{\mathcal{H}} \in \mathbb{R}^{\bar{J}_0 \cdots \times \bar{J}_{\mathrm{c}} \cdots \times \bar{J}_C}$ and $\bar{J}_{\mathrm{c}} \leq SI_{\mathrm{c}}$.

B. Derivation

For notational simplicity, we re-write the loss function as,

$$e := \|\mathcal{D} - \tilde{\mathcal{Z}}_{\mathcal{H}} \times_0 \tilde{\mathbf{U}}_{0\mathrm{x}} \ldots \times_{\mathrm{c}} \tilde{\mathbf{U}}_{\mathrm{cx}} \ldots \times_{\mathrm{C}} \tilde{\mathbf{U}}_{\mathrm{Cx}}\|^2 + \sum_{c=0}^{C} \sum_{s=1}^{S} \lambda_{\mathrm{c,s}} \|\tilde{\mathbf{U}}_{\mathrm{c,s}}^{\mathrm{T}} \tilde{\mathbf{U}}_{\mathrm{c,s}} - \mathbf{I}\| \quad (14)$$

where $\tilde{\mathbf{U}}_{\mathrm{cx}} = \mathbf{I}_{\mathrm{cx}} \bar{\mathbf{U}}_{\mathrm{c}\mathcal{H}} \tilde{\mathbf{G}}_{\mathrm{c}} = [\tilde{\mathbf{U}}_{\mathrm{c,1}} | \ldots | \tilde{\mathbf{U}}_{\mathrm{c,s}} | \ldots | \tilde{\mathbf{U}}_{\mathrm{c,S}}]$ and $\tilde{\mathbf{G}}_{\mathrm{c}} \in \mathbb{R}^{\bar{J}_c \times SI_c}$ is permutation matrix that groups the columns of $\mathbf{U}_{\mathrm{c}\mathcal{H}}$ based on the segment, s, to which they belong, and the inverse permutation matrices have been multiplied[3] into $\tilde{\mathcal{Z}}_{\mathcal{H}}$ resulting into a core that has also been grouped based on segments and sorted based on variance. The data tensor, $\mathcal{D}$, may be expressed in matrix form as in eq. 16 and reduces to the more efficiently block structure as in eq. 17

$$\mathcal{D} = \mathcal{Z}_{\mathcal{H}} \times_0 \mathbf{U}_{0\mathrm{x}} \times_1 \mathbf{U}_{1\mathrm{x}} \cdots \times_{\mathrm{c}} \mathbf{U}_{\mathrm{cx}} \cdots \times_{\mathrm{C}} \mathbf{U}_{\mathrm{Cx}} \quad (15)$$

$$\mathbf{D}_{[\mathrm{c}]} = \mathbf{U}_{\mathrm{cx}} \mathbf{Z}_{\mathcal{H}[\mathrm{c}]} (\mathbf{U}_{\mathrm{Cx}} \otimes \cdots \otimes \mathbf{U}_{(\mathrm{c}+1)\mathrm{x}} \otimes \mathbf{U}_{(\mathrm{c}-1)\mathrm{x}} \otimes \cdots \otimes \mathbf{U}_{0\mathrm{x}})^{\mathrm{T}} \quad (16)$$

$$= [\mathbf{U}_{\mathrm{c},1} \ldots \mathbf{U}_{\mathrm{c,s}} \ldots \mathbf{U}_{\mathrm{c,S}}] \quad (17)$$

$$\begin{bmatrix} \mathbf{Z}^{+}_{0[\mathrm{c}]} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \mathbf{Z}^{+}_{\mathrm{s}[\mathrm{c}]} & \mathbf{0} \\ & & \mathbf{0} & \ddots \\ \mathbf{0} & \cdots & & \mathbf{0} \quad \mathbf{Z}^{+}_{\mathrm{S}[\mathrm{c}]} \end{bmatrix} \begin{bmatrix} (\mathbf{U}_{\mathrm{C},1} \ldots \otimes \mathbf{U}_{(\mathrm{c}+1),1} \otimes \mathbf{U}_{(\mathrm{c}-1),1} \ldots \otimes \mathbf{U}_{0,1})^{\mathrm{T}} \\ \vdots \\ (\mathbf{U}_{\mathrm{C,s}} \ldots \otimes \mathbf{U}_{(\mathrm{c}+1),\mathrm{s}} \otimes \mathbf{U}_{(\mathrm{c}-1),\mathrm{s}} \ldots \otimes \mathbf{U}_{0,\mathrm{s}})^{\mathrm{T}} \\ \vdots \\ (\mathbf{U}_{\mathrm{C,S}} \ldots \otimes \mathbf{U}_{(\mathrm{c}+1),\mathrm{S}} \otimes \mathbf{U}_{(\mathrm{c}-1),\mathrm{S}} \ldots \otimes \mathbf{U}_{0,\mathrm{S}})^{\mathrm{T}} \end{bmatrix}$$

$$(\mathbf{U}_{\mathrm{Cx}} \cdots \odot \mathbf{U}_{(\mathrm{c}+1)\mathrm{x}} \odot \mathbf{U}_{(\mathrm{c}-1)\mathrm{x}} \cdots \odot \mathbf{U}_{0\mathrm{x}})^{\mathrm{T}}$$

$$= \mathbf{U}_{\mathrm{cx}} \mathbf{W}_{\mathrm{c}}^{\mathrm{T}}, \quad (18)$$

where $\otimes$ is the Kronecker product[6], and $\odot$ is the block-matrix Kahtri-Rao product.[7]

The matrixized block diagonal form of $\mathcal{Z}_{\mathcal{H}}$ in eq. 17 becomes evident when employing our modified data centric matrixizing operator based on the defintion 2, where the initial mode is the measurement mode.

The hierarchical block multilinear factorization, the M-mode Block SVD algorithm computes the mode matrix, $\mathbf{U}_{\mathrm{cx}}$, by computing the minimum of $e = \|\mathcal{D} - \tilde{\mathcal{Z}}_{\mathcal{H}} \times_0 \tilde{\mathbf{U}}_{0\mathrm{x}} \cdots \times_{\mathrm{C}} \tilde{\mathbf{U}}_{\mathrm{Cx}}\|^2$ by cycling through the modes, solving for $\tilde{\mathbf{U}}_{\mathrm{cx}}$ in the equation $\partial e / \partial \mathbf{U}_{\mathrm{cx}} = 0$ while holding the core tensor $\mathcal{Z}_{\mathcal{H}}$ and all the other mode matrices constant, and repeating until convergence. Note that

$$\frac{\partial e}{\partial \mathbf{U}_{\mathrm{cx}}} = \frac{\partial}{\partial \mathbf{U}_{\mathrm{cx}}} \|\mathbf{D}_{[\mathrm{c}]} - \mathbf{U}_{\mathrm{cx}} \mathbf{W}_{\mathrm{c}}^{\mathrm{T}}\|^2 = -\mathbf{D}_{[\mathrm{c}]} \mathbf{W}_{\mathrm{c}} + \mathbf{U}_{\mathrm{cx}} \mathbf{W}_{\mathrm{c}}^{\mathrm{T}} \mathbf{W}_{\mathrm{c}}.$$

Thus, $\partial e / \partial \mathbf{U}_{\mathrm{cx}} = 0$ implies that

$$\mathbf{U}_{\mathrm{cx}} = \mathbf{D}_{[\mathrm{c}]} \mathbf{W}_{\mathrm{c}} \left(\mathbf{W}_{\mathrm{c}}^{\mathrm{T}} \mathbf{W}_{\mathrm{c}}\right)^{-1} = \mathbf{D}_{[\mathrm{c}]} \mathbf{W}_{\mathrm{c}}^{\mathrm{T}+}$$

$$= \mathbf{D}_{[\mathrm{c}]} \left(\mathbf{Z}_{\mathcal{H}[\mathrm{c}]} (\mathbf{U}_{\mathrm{Cx}} \otimes \ldots \mathbf{U}_{(\mathrm{c}+1)\mathrm{x}} \otimes \mathbf{U}_{(\mathrm{c}-1)\mathrm{x}} \otimes \ldots \mathbf{U}_{0\mathrm{x}})^{\mathrm{T}}\right)^{+}$$

$$= \mathbf{D}_{[\mathrm{c}]} (\mathbf{U}_{\mathrm{Cx}} \odot \ldots \mathbf{U}_{(\mathrm{c}+1)\mathrm{x}} \odot \mathbf{U}_{(\mathrm{c}-1)\mathrm{x}} \odot \ldots \mathbf{U}_{0\mathrm{x}})^{\mathrm{T}+} \begin{bmatrix} \mathbf{Z}^{+}_{0[\mathrm{c}]} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \mathbf{Z}^{+}_{\mathrm{s}[\mathrm{c}]} & \mathbf{0} \\ & & \mathbf{0} & \ddots \\ \mathbf{0} & \cdots & & \mathbf{0} \quad \mathbf{Z}^{+}_{\mathrm{S}[\mathrm{c}]} \end{bmatrix}$$

whose $\mathbf{U}_{\mathrm{c,s}}$ sub-matrices are then subject to orthonormality constraints.

Solving for the optimal core tensor, $\mathcal{Z}_{\mathcal{H}}$, the data tensor, $\mathcal{D}$, approximation is expressed in vector form as,

$$e = \|\mathrm{vec}(\mathcal{D}) - (\tilde{\mathbf{U}}_{\mathrm{Cx}} \otimes \cdots \otimes \tilde{\mathbf{U}}_{\mathrm{cx}} \otimes \cdots \otimes \tilde{\mathbf{U}}_{0\mathrm{x}}) \mathrm{vec}(\tilde{\mathcal{Z}}_{\mathcal{H}})\|. \quad (19)$$

[3] Equivalent representations can be transformed into one another by post-multiplying mode matrices with nonsingular matrices, $\mathbf{G}_{\mathrm{c}}$, $\mathcal{D} = (\mathcal{Z}_{\mathcal{H}} \times_0 \mathbf{G}_0^{-1} \cdots \times_{\mathrm{c}} \mathbf{G}_{\mathrm{c}}^{-1} \cdots \times_{\mathrm{C}} \mathbf{G}_{\mathrm{C}}^{-1}) \times_1 \mathbf{I}_{1\mathrm{x}} \mathbf{U}_{1\mathcal{H}} \mathbf{G}_1 \cdots \times_{\mathrm{c}} \mathbf{I}_{\mathrm{cx}} \mathbf{U}_{\mathrm{c}\mathcal{H}} \mathbf{G}_{\mathrm{c}} \cdots \times_{\mathrm{C}} \mathbf{I}_{\mathrm{Cx}} \mathbf{U}_{\mathrm{C}\mathcal{H}} \mathbf{G}_{\mathrm{C}}$.

[4] Note that eq.(13) does not reduce to a multilinear subspace decomposition of $\mathcal{D}_{\mathcal{H}}$ since $\mathcal{D} \times_0 \mathbf{I}_{0\mathrm{x}}^{+} \times_1 \mathbf{I}_{1\mathrm{x}}^{+} \cdots \times_{\mathrm{c}} \mathbf{I}_{\mathrm{cx}}^{+} \cdots \times_{\mathrm{C}} \mathbf{I}_{\mathrm{Cx}}^{+} \neq \mathcal{D}_{\mathcal{H}}$.

[5] For computational efficiency, we may perform M-mode SVD on each data tensor segment $\mathcal{D}_{\mathrm{s}}$ and concatenate terms along the diagonal of $\mathcal{Z}_{\mathcal{H}}$ and $\mathbf{U}_{\mathrm{c}\mathcal{H}}$.

[6] The Kronecker product of $\mathbf{U} \in \mathbb{R}^{I \times J}$ and $\mathbf{V} \in \mathbb{R}^{K \times L}$ is the $IK \times JL$ matrix defined as $[\mathbf{U} \otimes \mathbf{V}]_{ik,jl} = u_{ij} v_{kl}$.

[7] The Khatri-Rao product of $[\mathbf{U}_1 \ldots \mathbf{U}_{\mathrm{n}} \ldots \mathbf{U}_{\mathrm{N}}] \odot [\mathbf{V}_1 \ldots \mathbf{V}_{\mathrm{n}} \ldots \mathbf{V}_{\mathrm{N}}]$ with $\mathbf{U}_{\mathrm{l}} \in \mathbb{R}^{I \times N_{\mathrm{l}}}$ and $\mathbf{V}_{\mathrm{l}} \in \mathbb{R}^{K \times N_{\mathrm{l}}}$ is a *block-matrix Kronecker product*; therefore, it can be expressed as $\mathbf{U} \odot \mathbf{V} = [(\mathbf{U}_1 \otimes \mathbf{V}_1) \ldots (\mathbf{U}_{(\mathrm{l})} \otimes \mathbf{V}_{(\mathrm{l})}) \ldots (\mathbf{U}_{(\mathrm{L})} \otimes \mathbf{V}_{(\mathrm{L})})]$.

Algorithm 1 M-mode Block SVD.

Input: Data tensor, $\mathcal{D} \in \mathbb{R}^{I_0 \times I_1 \times \cdots \times I_C}$, filters $\mathbf{H}_s$, and desired dimensionality reduction $\tilde{J}_1, \ldots, \tilde{J}_C$.

1.*Initialization:*

1a. Decompose each data tensor segment, $\mathcal{D}_s = \mathcal{D} \times \mathbf{H}_s$, by employing the M-mode SVD.

$$\mathcal{D}_s = \mathcal{Z} \times_0 \mathbf{U}_{0,s} \cdots \times_c \mathbf{U}_{c,s} \cdots \times_C \mathbf{U}_{C,s}$$

1b. For $c = 0, 1, \ldots, C$, set $\mathbf{U}_{cx} = [\mathbf{U}_{c,1} \ldots \mathbf{U}_{c,s} \ldots \mathbf{U}_{c,S}]$, and truncate to $\tilde{J}_c$ columns by sorting all the eigenvalues from all data segments and deleting the columns corresponding to the lowest eigenvalues from various $\mathbf{U}_{c,s}$ and the rows from $\mathbf{Z}_{s[c]}$

2.*Optimization via alternating least squares:*

Iterate for $n := 1, \ldots, N$

For $c := 0, \ldots, C$,

2a. Compute mode matrix $\mathbf{U}_{cx}$ while holding the rest fixed.

$$\mathbf{U}_{cx} := \mathbf{D}_{[c]} \left(\mathbf{U}_{Cx} \odot \ldots \mathbf{U}_{(c+1)x} \odot \mathbf{U}_{(c-1)x} \odot \ldots \mathbf{U}_{0x}\right)^{\mathrm{T}} \begin{bmatrix} \mathbf{Z}^{+}_{0[c]} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \mathbf{Z}^{+}_{s[c]} & \mathbf{0} \\ & & \mathbf{0} & \ddots \; \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Z}^{+}_{S[c]} \end{bmatrix}$$

Set $\hat{\mathbf{U}}_{c,s}$ to the $\tilde{R}_{c,s}$ leading left-singular vectors of the SVD[a] of $\mathbf{U}_{c,s}$, a subset of the columns in $\mathbf{U}_{cx}$.Update $\mathbf{U}_{c,s}$ in $\mathbf{U}_{cx}$ with $\hat{\mathbf{U}}_{c,s}$

2b. Set the non-zero(nz) entries of $\mathrm{vec}(\mathcal{Z}_{\mathcal{H}})_{\mathrm{nz}}$ based on:

$$\mathrm{vec}(\mathcal{Z}_{\mathcal{H}})_{\mathrm{nz}} = (\mathbf{U}_{Cx} \otimes \cdots \otimes \mathbf{U}_{cx} \otimes \cdots \otimes \mathbf{U}_{0x})^{+}_{\mathrm{nz}} \mathrm{vec}(\mathcal{D}).$$

until convergence. [b]

Output converged matrices $\mathbf{U}_{1x}, \ldots, \mathbf{U}_{Cx}$ and tensor $\mathcal{Z}_{\mathcal{H}}$.

[a] The complexity of computing the SVD of an $m \times n$ matrix $\mathbf{A}$ is $O(mn \min(m,n))$, which is costly when both m and n are large. However, we can efficiently compute the $\tilde{R}$ leading left-singular vectors of $\mathbf{A}$ by first computing the rank-$\tilde{R}$ modified Gram Schmidt (MGS) orthogonal decomposition $\mathbf{A} \approx \mathbf{QR}$, where $\mathbf{Q}$ is $m \times \tilde{R}$ and $\mathbf{R}$ is $\tilde{R} \times n$, and then computing the SVD of $\mathbf{R}$ and multiplying it as follows: $\mathbf{A} \approx \mathbf{Q}(\tilde{\mathbf{U}}\mathbf{S}\mathbf{V}^{\mathrm{T}}) = (\mathbf{Q}\tilde{\mathbf{U}})\mathbf{S}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$.

[b] Note that N is a pre-specified maximum number of iterations. A possible convergence criterion is to compute at each iteration the approximation error $e_n := \|\mathcal{D} - \hat{\mathcal{D}}\|^2$ and test if $e_{n-1} - e_n \leq \epsilon$ for sufficiently small tolerance ϵ.

Solve for the non-zero(nz) terms of $\mathcal{Z}_{\mathcal{H}}$ in the equation $\partial e / \partial(\mathcal{Z}_{\mathcal{H}}) = 0$, by removing the corresponding zero columns of the first matrix on right side of the equation below, performing the pseudo-inverse, and setting

$$\mathrm{vec}(\mathcal{Z}_{\mathcal{H}})_{\mathrm{nz}} = (\mathbf{U}_{Cx} \otimes \cdots \otimes \mathbf{U}_{cx} \otimes \cdots \otimes \mathbf{U}_{0x})^{+}_{\mathrm{nz}} \mathrm{vec}(\mathcal{D}). \quad (20)$$

Repeat all steps until convergence.This optimization is the basis of the M-mode Block SVD Algorithm 1.

When the data tensor is a collection of observations made up of non-overlapping parts, Fig. 3d, the data tensor decomposition reduces to the concatenation of an M-mode SVD of individual parts and when the data tensor is a collection of overlapping parts that have the same multilinear-rank reduction, Fig.3e, see[42] for additional specific optimizations.

IV. Representing Levels of Abstraction Bottom-up

An incremental hierarchical block multilinear factorization that represents levels of abstractions bottom-up is developed analogously to the incremental SVD for matrices [4]. The precomputed multilinear factorizations of the children parts are employed to determine the parent whole multilinear factorization. The derived algorithm may also be employed to update the overall model when the data becomes available sequentially [23]. We first address the computation of the mode matrices and the extended core of the parent whole when the children parts are non-overlapping. Next, we consider the overlapping children case, and the case where the parent-wholes and children-parts contain differently filtered data.

Computing parent causal mode matrices, $\mathbf{U}_{c,w}$: Note that the parent whole, $\mathcal{D}_w$, is a concatenation of the data contained in its K' children segments that are part of the hierarchy, $\mathcal{D}_k$, where $1 \leq k \leq K'$. New data that is not contained by any of the children is denoted as the $K = K' + 1$ child,$\mathcal{D}_{\mathrm{K}}$, eq. 21 . We initialize the hierarchical block multilinear factorization by performing an M-mode SVD on each leaf.

The c mode matrix, $\mathbf{U}_{c,w}$ of the w parent whole, $\mathcal{D}_w$, is the left singular matrix of $[\, \mathbf{U}_{c,1}\mathbf{\Sigma}_{c,1} \quad \ldots \quad \mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k} \quad \ldots \quad \mathbf{U}_{c,K}\mathbf{\Sigma}_{c,K} \,]$ which is based on the following derivation, that writes SVD of the flattened parent whole in terms of the SVDs of its flattened children parts, followed by a collection terms such that $\mathbf{V}_{c,\mathrm{all}}$ is a block diagonal matrix of $\mathbf{V}_{c,k}$:

$$\begin{aligned} \mathbf{D}_{w[c]} &= [\, \mathbf{D}_{1[c]} \quad \cdots \quad \mathbf{D}_{k[c]} \quad \cdots \quad \mathbf{D}_{K[c]} \,] && (21) \\ &= [\, \mathbf{U}_{c,1}\mathbf{\Sigma}_{c,1}\mathbf{V}^{\mathrm{T}}_{c,1} \quad \cdots \quad \mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k}\mathbf{V}^{\mathrm{T}}_{c,k} \quad \cdots \quad \mathbf{U}_{c,K}\mathbf{\Sigma}_{c,K}\mathbf{V}^{\mathrm{T}}_{c,K} \,] \\ &= \underbrace{[\, \mathbf{U}_{c,1}\mathbf{\Sigma}_{c,1} \quad \ldots \quad \mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k} \quad \ldots \quad \mathbf{U}_{c,K}\mathbf{\Sigma}_{c,K} \,]}_{\text{QR + SVD of R}} \mathbf{V}^{\mathrm{T}}_{c,\mathrm{all}} && (22) \\ &= \mathbf{U}_{c,w}\mathbf{\Sigma}_{c,w} \underbrace{[\, \mathbf{V}^{\mathrm{T}}_{c,w1} \quad \ldots \quad \mathbf{V}^{\mathrm{T}}_{c,wk} \quad \ldots \quad \mathbf{V}^{\mathrm{T}}_{c,wK} \,]}_{\mathbf{V}^{\mathrm{T}}_{c,w}} \mathbf{V}^{\mathrm{T}}_{c,\mathrm{all}}, && (23) \end{aligned}$$

Computing the parent extended core, $\mathcal{T}_w$: Computation of the extended core associated with the parent whole, $\mathcal{T}_w$, is performed by considering the following derivation

$$\begin{aligned} \mathbf{D}_{w[c]} &= [\, \mathbf{D}_{1[c]} \quad \ldots \quad \mathbf{D}_{k[c]} \quad \ldots \quad \mathbf{D}_{K[c]} \,] && (24) \\ &= \big[\, \mathbf{U}_{c,1}\mathbf{\Sigma}_{c,1}\hat{\mathbf{T}}_{1[c]}(\mathbf{U}_{1,1} \otimes \ldots \mathbf{U}_{c\text{-}1,1} \otimes \mathbf{U}_{c+1,1} \otimes \ldots \mathbf{U}_{C,1})^{\mathrm{T}} \quad \cdots \quad \mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k}\hat{\mathbf{T}}_{k[c]}(\mathbf{U}_{c\text{-}1,0} \otimes \ldots \mathbf{U}_{c\text{-}1,k} \otimes \mathbf{U}_{c+1,k} \otimes \ldots \mathbf{U}_{C,k})^{\mathrm{T}} \quad \cdots \quad \big] \\ &= \underbrace{[\, \mathbf{U}_{c,1}\mathbf{\Sigma}_{c,1} \; \ldots \; \mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k} \; \ldots \; \mathbf{U}_{c,K}\mathbf{\Sigma}_{c,K} \,]}_{\text{QR + SVD of R}} \begin{bmatrix} \hat{\mathbf{T}}_{1[c]}(\mathbf{U}_{1,1} \otimes \ldots \mathbf{U}_{c\text{-}1,1} \otimes \mathbf{U}_{c+1,1} \otimes \ldots \mathbf{U}_{C,1})^{\mathrm{T}} & \mathbf{0} & \cdots \\ & & \vdots \\ \mathbf{0} & \ddots & \mathbf{0} \\ \vdots \ldots & \mathbf{0} & \hat{\mathbf{T}}_{K[c]}(\mathbf{U}_{1,K} \otimes \ldots \mathbf{U}_{c\text{-}1,K} \otimes \mathbf{U}_{c+1,K} \otimes \ldots \mathbf{U}_{C,K})^{\mathrm{T}} \end{bmatrix} \\ &= \mathbf{U}_{c,w}\mathbf{\Sigma}_{c,w} \underbrace{[\, \mathbf{V}^{\mathrm{T}}_{c,1} \quad \ldots \quad \mathbf{V}^{\mathrm{T}}_{c,k} \quad \ldots \quad \mathbf{V}^{\mathrm{T}}_{c,K} \,]}_{\mathbf{V}^{\mathrm{T}}_{c}} \begin{bmatrix} \hat{\mathbf{T}}_{1[c]}(\mathbf{U}_{1,1} \otimes \ldots \mathbf{U}_{c\text{-}1,1} \otimes \mathbf{U}_{c+1,1} \otimes \ldots \mathbf{U}_{C,1})^{\mathrm{T}} & \mathbf{0} & \cdots \\ & & \vdots \\ \mathbf{0} & \ddots & \mathbf{0} \\ \vdots \ldots & \mathbf{0} & \hat{\mathbf{T}}_{K[c]}(\mathbf{U}_{1,K} \otimes \ldots \mathbf{U}_{c\text{-}1,K} \otimes \mathbf{U}_{c+1,K} \otimes \ldots \mathbf{U}_{C,K})^{\mathrm{T}} \end{bmatrix}. \end{aligned}$$

Algorithm 2 Incremental M-Mode Block SVD.

Input the data tensor $\mathcal{D} \in \mathbb{R}^{I_0 \times \cdots \times I_M}$ and data tensor tree parameterization.

1) For $c := 0, \ldots, C$, in a bottom-up fashion

1a. For each data tensor leaf nodes:
Set mode matrix $\mathbf{U}_{c,s}$ to the left matrix of the SVD of $\mathbf{D}_{s[c]}\mathbf{D}_{s[c]}^{\mathrm{T}} = \mathbf{U}_{c,s}\mathbf{\Sigma}_{c,s}^2\mathbf{U}_{c,s}^{\mathrm{T}}$. [a]

1b. In a bottom-up fashion, for each data tensor that is a parent-whole:
Set the mode matrix $\mathbf{U}_{c,s}$ of the parent-whole by computing the SVD

$$\mathbf{U}_{c,w}\mathbf{\Sigma}_{c,w}\mathbf{V}_{c,w}^{\mathrm{T}} = \mathrm{SVD}([\,\mathbf{U}_{c,1}\mathbf{\Sigma}_{c,1} \quad \ldots \quad \mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k} \quad \ldots \quad \mathbf{U}_{c,K}\mathbf{\Sigma}_{c,K} \quad \mathbf{U}_{c,n}\mathbf{\Sigma}_{c,n}\,]),$$ [a]

where $\mathbf{U}_{c,k}\mathbf{\Sigma}_{c,k}$ in the mode matrix and singular value matrix for the c causal factor of the k child, for $1 \leq k \leq K$.

2) Compute the normalized core of the k child $\hat{\mathcal{T}}_k = \mathcal{T}_k \times_1 \mathbf{\Sigma}_{1,k}^{-1} \cdots \times_c \mathbf{\Sigma}_{c,k}^{-1} \cdots \times_C \mathbf{\Sigma}_{C,k}^{-1}$

Let $\hat{\mathcal{T}}_{\text{Kall}}$ be a tensor that contains the children normalized cores $\hat{\mathcal{Z}}_k$ along its super-diagonal.
Set the core of the parent whole as $\mathcal{T}_w = \hat{\mathcal{T}}_{\text{Kall}} \times_1 \mathbf{\Sigma}_{1,w}\mathbf{V}_1^{\mathrm{T}} \cdots \times_c \mathbf{\Sigma}_{c,w}\mathbf{V}_c^{\mathrm{T}} \cdots \times_C \mathbf{\Sigma}_{C,w}\mathbf{V}_{C,w}^{\mathrm{T}}$.

Output mode matrices $\mathbf{U}_0, \ldots, \mathbf{U}_C$ and the core tensor $\mathcal{T}$ of the final parent-whole.

[a] The complexity of computing the SVD of an $m \times n$ matrix $\mathbf{A}$ is $O(mn\min(m,n))$, which is costly when both m and n are large. However, we can efficiently compute the $\tilde{R}$ leading left-singular vectors of $\mathbf{A}$ by first computing the rank-$\tilde{R}$ modified Gram-Schmidt (MGS) orthogonal decomposition $\mathbf{A} \approx \mathbf{QR}$, where $\mathbf{Q}$ is $m \times \tilde{R}$ and $\mathbf{R}$ is $\tilde{R} \times n$, and then computing the SVD of $\mathbf{R}$ and multiplying it as follows: $\mathbf{A} \approx \mathbf{Q}(\tilde{\mathbf{U}}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}) = (\mathbf{Q}\tilde{\mathbf{U}})\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$.

where $\hat{\mathbf{T}}_{\text{k[c]}} = \mathbf{\Sigma}_{\text{c,k}}^{-1}\mathbf{T}_{\text{k[c]}}$. Let $\hat{\mathcal{T}}_{\text{k}} = \mathcal{T}_{\text{k}} \times_1 \mathbf{\Sigma}_{\text{1,k}}^{-1} \cdots \times_{\text{c}} \mathbf{\Sigma}_{\text{c,k}}^{-1} \cdots \times_{\text{C}} \mathbf{\Sigma}_{\text{C,k}}^{-1}$ is the normalized extended core of the k^{th} child, and $\hat{\mathcal{T}}_{\text{Kall}}$ contains along the diagonal the children normalized extended cores $\hat{\mathcal{T}}_{\text{k}}$. Thus, the extended core of the parent whole is

$$\mathcal{T}_{\text{w}} = \hat{\mathcal{T}}_{\text{Kall}} \times_1 \mathbf{\Sigma}_{\text{1,w}}\mathbf{V}_1^{\mathrm{T}} \cdots \times_{\text{c}} \mathbf{\Sigma}_{\text{c,w}}\mathbf{V}_{\text{c}}^{\mathrm{T}} \cdots \times_{\text{C}} \mathbf{\Sigma}_{\text{C,w}}\mathbf{V}_{\text{C,w}}^{\mathrm{T}}. \quad (25)$$

Overlapping children: This case may be reduced to the non-overlapping case by introducing another level in the hierarchy. Overlapping children are now treated as parents with one non-overlaping child sub-part and child sub-parts that correspond to every possible combination of overlaps that are shared by siblings. The original parent whole representation is computed in terms of the grandchildren representations.

Parent-whole and children-parts with differently filtered data: This is the case when a parent-whole and the children parts contain differently filtered information, as in the case when a parent-whole and the children parts sample information from different layers of a Laplacian pyramid. This case may be reduced to a non-overlapping case by writing the filters as the product between a segmentation filter, $\mathbf{S}$, *i.e.*, an identity matrix with limited spatial scope, and general filter that post multiplies the segmentation filter, $\mathbf{H}_{\text{s}} = \mathbf{F}_{\text{s}}\mathbf{S}_{\text{s}}$ and $\mathcal{D}_{\text{s}} = (\mathcal{D} \times_0 \mathbf{S}_{\text{s}}) \times_0 \mathbf{F}_{\text{s}}$. The general filters, $\mathbf{F}_{\text{s}}$ may be applied after the cores are computed.

Computational Cost Analysis: Let an M-order data tensor, $\mathcal{D} \in \mathbb{R}^{I_0 \times I_1 \cdots \times I_{\text{c}} \times \cdots \times I_{\text{C}}}$, where $M = C + 1$, be recursively subdivided into $K = 2^M$ children of the same order, but with each mode half in size. There are a total of $log_K N + 1$ levels, where $N = \prod_{i=0}^{C} I_i$. Recursive subdivision results in $S = N \log_{2^M} N + 1$ segments. The total computational cost is the amortized M-mode SVD cost per data tensor segment, T, times the number of segments, $\mathcal{O}(TN \log_K N)$. Since siblings at each level can be computed independently, on a distributed system the cost is $O(Tlog_K N)$.

V. CausalX Experiments

CausalX visual recognition system computes a set of causal explanations based on a counterfactual causal model that takes advantage of the assets of multilinear (tensor) algebra. The M-mode Block SVD and the Incremental M-mode Block SVD algorithms estimate the model parameters. In the context of face image verification, we compute a compositional hierarchical person representation [42]. Our system is trained on a set of observations that are the result of combinatorially manipulating the scene structure, the viewing and illumination conditions. We rendered in Maya images of 100 people from 15 different viewpoints with 15 different illuminations. The collection of vectorized images with $10,414$ pixels is organized in a data tensor, $\mathcal{D} \in \mathbb{R}^{10,414 \times 15 \times 15 \times 100}$. The counterfactual model is estimated by employing $\mathcal{D}_{\mathcal{H}}$, a hierarchical tensor of part-based Laplacian pyramids. We report encouraging face verification results on two test data sets the Freiburg, and the Labeled Faces in the Wild (LFW) datasets. We have currently achieved verification rates just shy of 80% on LFW [42], by employing less than one percent (1%) of the total images employed by DeepFace [35]. When data is limited, convolutional neural networks (CNNs) do not convergence or generalize. More importantly, CNNs are predictive rather than causal models.

Conclusion

This paper deepens the definition of causality in a multilinear (tensor) framework by addressing the distinctions between intrinsic versus extrinsic causality, and local versus global causality. It proposes a unified multilinear model of wholes and parts that reconceptualizes a data tensor in terms of a *hierarchical data tensor*. Our hierarchical data tensor is a mathematical instantiation of a tree data structure that enables a single elegant model of wholes and parts and allows for different tree parameterizations for the intrinsic versus extrinsic causal factors. The derived tensor factorization is a hierarchical block multilinear factorization that disentangles the causal structure of data formation. Given computational efficiency considerations, we present an incremental computational alternative that employs the part representations from the lower levels of abstraction to compute the parent whole representations from the higher levels of abstraction in an iterative bottom-up way. This computational approach may be employed to update causal representations in scenarios when data is available incrementally. The resulting object representation is a combinatorial choice of part representations, that renders object recognition robust to occlusion and reduces large training data requirements. We have demonstrated our work in the context of face verification by extending the TensorFaces method with promising results. TensorFaces is a component of CausalX, a counterfactual causal based visual recognition system, and an explainable AI.

Acknowledgement

The authors are thankful to Ernest Davis for feedback provided during the writing of this document, and to Donald Rubin and Andrew Gelman for helpful discussions.

REFERENCES

[1] E. Acar, E. E. Papalexakis, G. Gürdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, and R. Bro. Structure-revealing data fusion. BMC bioinformatics, 15(1):239, 2014.

[2] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. Proc. of the National Academy of Sciences, 113(27):7345–52, 2016.

[3] P. M. Bentler and S.-Y. Lee. A statistical development of three-mode factor analysis. British J. of Math. and Stat. Psych., 32(1):87–104, 1979.

[4] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In Proc. 7th European Conf. on Computer Vision (ECCV), volume 2350, pages 707–20. Springer, May 2002.

[5] R. Bro. Parafac: Tutorial and applications. In Chemom. Intell. Lab Syst., Special Issue 2nd Internet Cont. in Chemometrics (INCINC'96), volume 38, pages 149–171, 1997.

[6] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. Psychometrika, 35:283–319, 1970.

[7] W. Chu and Z. Ghahramani. Probabilistic models for incomplete multi-dimensional arrays. volume 5 of Proceedings of Machine Learning Research, pages 89–96, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[8] W. G. Cochran. Observational studies. In T. Bancroft, editor, Statistical Papers in Honor of George W. Snedecor, pages 77–90. Iowa State University Press, 1972.

[9] L. de Lathauwer. Signal Processing Based on Multilinear Algebra. PhD thesis, Katholieke Univ. Leuven, Belgium, 1997.

[10] L. de Lathauwer. Decompositions of a higher-order tensor in block termspart ii: Definitions and uniqueness. SIAM J. on Matrix Analysis and Applications, 30(3):1033–1066, 2008.

[11] A. Elgammal and C. S. Lee. Separating style and content on a nonlinear manifold. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), volume I, pages 478–485, Jun 2004.

[12] A. Gelman and G. Imbens. Why ask why? forward causal inference and reverse causal questions. Tech.report, Nat.Bureau of Econ Research, 2013.

[13] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th Inter. Conf. on Data Science and Advanced Analytics (DSAA), pages 80–89. IEEE, 2018.

[14] C. Glymour. Statistics and causal inference: Comment: Statistics and metaphysics. J. of the American Stat. Assoc., 81(396):964–66, Dec 1986.

[15] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a right to explanation. AI Magazine, 38(3):50–57, Oct. 2017.

[16] R. Harshman. Foundations of the PARAFAC procedure: Model and conditions for an explanatory factor analysis. Tech. Report Working Papers in Phonetics 16, UCLA, CA, Dec 1970.

[17] P. W. Holland. Statistics and causal inference: Rejoinder. J. of the American Statistical Association, 81(396):968970, 1986.

[18] E. Hsu, K. Pulli, and J. Popovic. Style translation for human motion. ACM Transactions on Graphics, 24(3):1082–89, 2005.

[19] I. Humberstone. Intrinsi/extrinsic. Synthese, 108:206–267, 1986.

[20] G. W. Imbens and D. B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge Univ. Press, 2015.

[21] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. SIAM review, 51(3):455–500, 2009.

[22] D. Lewis. Extrinsic properties. Philosophical Studies, 44:197–200, 1983.

[23] X. Li, W. Hu, Z. Zhang, X. Zhang, and G. Luo. Robust visual tracking based on incremental tensor subspace learning. In 2007 IEEE 11th Inter. Conf. on Computer Vision, pages 1–8, 2007.

[24] L. Lim and P. Comon. Blind multilinear identification. IEEE Transactions on Information Theory, 60(2):1260–1280, 2014.

[25] G. Liu, M. Xu, Z. Pan, and A. E. Rhalibi. Human motion generation with multifactor models. Computer Animation and Virtual Worlds, 22(4):351–359, 2011.

[26] J. R. Magnus and H. Neudecker. Matrix Differential Calculus with Applications in Statistics and Econometrics. John Wiley & Sons, 1988.

[27] P. P. Markopoulos, D. G. Chachlakis, and A. Prater-Bennette. L1-norm higher-order singular-value decomposition. In 2018 IEEE Global Conf. on Signal and Information Processing (GlobalSIP), pages 1353–1357, 2018.

[28] E. Miandji, S. Hajisharif, and J. Unger. A unified framework for compression and compressed sensing of light fields and light field videos. ACM Trans. Graph., 38(3), May 2019.

[29] J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge Univ. Press, 2000.

[30] J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. Statistical Science, 29(4):57995, 2014.

[31] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. J. of Educational Psych., 66(5):688–701, 1974.

[32] D. B. Rubin. Bayesian inference for causality: The importance of randomization. In The Proceedings of the Social Statistics Section. 1975.

[33] N. D. Sidiropoulos, L. de Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. IEEE Transactions on Signal Processing, 65:3551–82, 2017.

[34] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. Causation, prediction, and search. MIT press, 2000.

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pages 1701–08, 2014.

[36] Y. Tang, R. Salakhutdinov, and G. Hinton. Tensor analyzers. volume 28 of Proceedings of Machine Learning Research, pages 163–171, Atlanta, Georgia, USA, 17–19 Jun 2013.

[37] L. R. Tucker. Some mathematical notes on three-mode factor analysis. Psychometrika, 31:279–311, 1966.

[38] M. Vasilescu and D. Terzopoulos. Adaptive meshes and shells: Irregular triangulation, discontinuities, and hierarchical subdivision. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'92), page 829832, Champaign, IL, Jun 1992.

[39] M. A. O. Vasilescu. Human motion signatures: Analysis, synthesis, recognition. In Proc. Int. Conf. on Pattern Recognition, volume 3, pages 456–460, Quebec City, Aug 2002.

[40] M. A. O. Vasilescu. A Multilinear (Tensor) Algebraic Framework for Computer Graphics, Computer Vision, and Machine Learning. PhD thesis, University of Toronto, 2009.

[41] M. A. O. Vasilescu. Multilinear projection for face recognition via canonical decomposition. In Proc. IEEE Inter. Conf. on Automatic Face Gesture Recognition (FG 2011), pages 476–483, Mar 2011.

[42] M. A. O. Vasilescu and E. Kim. Compositional hierarchical tensor factorization: Representing hierarchical intrinsic and extrinsic causal factors. In The 25th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD19): Tensor Methods for Emerging Data Science Challenges Workshop, Aug. 5 2019.

[43] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In Proc. European Conf. on Computer Vision (ECCV 2002), pages 447–460, Copenhagen, Denmark, May 2002.

[44] M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, volume II, pages 93–99, Madison, WI, 2003.

[45] M. A. O. Vasilescu and D. Terzopoulos. TensorTextures: Multilinear image-based rendering. ACM Transactions on Graphics, 23(3):336–342, Aug 2004. Proc. ACM SIGGRAPH 2004 Conf., Los Angeles, CA.

[46] M. A. O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, volume I, pages 547–553, San Diego, CA, 2005.

[47] M. A. O. Vasilescu and D. Terzopoulos. Multilinear projection for appearance-based recognition in the tensor framework. In Proc. 11th IEEE Inter. Conf. on Computer Vision (ICCV'07), pages 1–8, 2007.

[48] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. ACM Transactions on Graphics (TOG), 24(3):426–433, Jul 2005.

[49] H. Wang and N. Ahuja. Facial expression decomposition. In Proc, 9th IEEE Inter. Conf. on Computer Vision (ICCV), pages 958–65,v.2, 2003.

[50] M. Wang, Y. Panagakis, P. Snape, and S. Zafeiriou. Learning the multilinear structure of visual data. In 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 6053–6061, Jul 2017.