Reduced-Rank Tensor-on-Tensor Regression and Tensor-Variate Analysis of Variance

Carlos Llosa-Vite and Ranjan Maitra

Abstract—Fitting regression models with many multivariate responses and covariates can be challenging, but such responses and covariates sometimes have tensor-variate structure. We extend the classical multivariate regression model to exploit such structure in two ways: first, we impose four types of low-rank tensor formats on the regression coefficients. Second, we model the errors using the tensor-variate normal distribution that imposes a Kronecker separable format on the covariance matrix. We obtain maximum likelihood estimators via block-relaxation algorithms and derive their computational complexity and asymptotic distributions. Our regression framework enables us to formulate tensor-variate analysis of variance (TANOVA) methodology. This methodology, when applied in a one-way TANOVA layout, enables us to identify cerebral regions significantly associated with the interaction of suicide attempters or non-attemptor ideators and positive-, negative- or death-connoting words in a functional Magnetic Resonance Imaging study. Another application uses three-way TANOVA on the Labeled Faces in the Wild image dataset to distinguish facial characteristics related to ethnic origin, age group and gender. A R package `totr` implements the methodology.

Index Terms—CP decomposition, HOLQ, HOSVD, kronecker separable models, LFW dataset, multilinear statistics, multiway regression, random tensors, suicide ideation, tensor train format, tensor ring format, tucker format

1 Introduction

The classical simple linear regression (SLR) model (without intercept) relates the response variable y_i to the explanatory variable x_i as $y_i = \beta x_i + e_i$ with $\mathbb{V}\mathrm{ar}(e_i) = \sigma^2$ for $i = 1, 2, \ldots, n$, where β is the regression coefficient parameter and σ^2 is the variance parameter. A natural extension of SLR for vector-valued responses and explanatory variables is the multivariate multiple linear regression (MVMLR) model

$$\boldsymbol{y}_i = B\boldsymbol{x}_i + \boldsymbol{e}_i, \quad \mathbb{V}\mathrm{ar}(\boldsymbol{e}_i) = \Sigma, \tag{1}$$

where $(\boldsymbol{y}_1, \boldsymbol{x}_1), (\boldsymbol{y}_2, \boldsymbol{x}_2), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$ are vector-valued responses and covariates and (B, Σ) are parameters. The number of parameters relative to the sample size in (1) is greater in the MVMLR model that in its SLR counterpart because the parameters (B, Σ) are matrix-valued [1]. Several methodologies, for example, the lasso and graphical lasso [2], [3], envelope models [4] and reduced rank regression [5], have been proposed to alleviate issues arising from the large number of parameters in (1). However, these methodologies are only for vector-valued observations and do not exploit their underlying structure that may further reduce the number of necessary parameters, in some cases making computation feasible. Here we consider *tensor-* or array-structured responses and covariates that arise in many applications, such as the two motivating examples introduced next.

1.1 Illustrative Examples

1.1.1 Cerebral Activity in Subjects at Risk of Suicide

The United States Center for Disease Control and Prevention reports that 47,173 Americans died by suicide in 2017, accounting for about two-thirds of all homicides in that year. Accurate suicide risk assessment is challenging, even for trained mental health professionals, as 78% of patients who commit suicide deny such ideation even in their last communication with professionals [6]. Understanding how subjects at risk of suicide respond to different stimuli is important to guide treatment and therapy. [7] provided data from a functional Magnetic Resonance Imaging (fMRI) study of nine suicide attempters and eight suicide non-attempter ideators (henceforth, ideators) upon exposing them to ten words each with positive, negative or death-related connotations. Our interest is in understanding brain regions associated with the interaction of a subject's attempter/ideator status and word type to inform diagnosis and treatment.

Traditional approaches fit separate regression models at each voxel without regard to spatial context that is only addressed post hoc at the time of inference. A more holistic strategy would use (1) with the response vector $\boldsymbol{y}_i$ of size $30 \times 43 \times 56 \times 20 = 1444800$, which contains thirty $43 \times 56 \times 20$ image volumes, for each of the $i = 1, 2, \ldots, 17$ subjects. The explanatory variable here is a 2D vector that indicates a subject's status as a suicide attempter or ideator. Under this

• *The authors are with the Department of Statistics, Iowa State University, Ames, IA 50011 USA. E-mail: {cllosa, maitra}@iastate.edu.*

Manuscript received 28 Mar. 2021; revised 3 Mar. 2022; accepted 28 Mar. 2022. Date of publication 5 Apr. 2022; date of current version 6 Jan. 2023.
This work was supported in part by the National Institute of Justice (NIJ) under Grants 2015-DN-BX-K056, 2018-R2-CX-0034, and 15PNIJ-21-GG-04141-RESS. The work of Ranjan Maitra was also supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH) under Grant R21EB016212, and in part by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Hatch project under Grant IOW03717.
(Corresponding author: Ranjan Maitra)
Recommended for acceptance by B. Hammer.
Digital Object Identifier no. 10.1109/TPAMI.2022.3164836

framework, B and Σ have over 2.8 million and 1 trillion unconstrained parameters, making estimation with only 17 subjects impractical. Incorporating a 3D spatial autoregressive (AR) structure into the image volume, and another correlation structure between the words can allow estimation of the variance, but still needs additional methodology to accomodate the large sixth-order tensor-structured regression parameter B. We develop such methodology in this paper, and return to this dataset in Section 4.1.

1.1.2 Distinguishing Characteristics of Faces

Distinguishing the visual characteristics of faces is important for biometrics. The Labeled Faces in the Wild (LFW) database [8] is used for developing and testing facial recognition methods, and contains over 13000 250×250 color images of faces of different individuals along with their classification into ethnic origin, age group and gender [9], [10]. We use a subset, totalling 605 images, for which the three attributes of ethnic origin (African, European or Asian, as specified in the database), cohort (child, youth, middle-aged or senior) and gender (male or female) are unambiguous. The color at each pixel is a 3D RGB vector so each image (response) is a $250 \times 250 \times 3$ array. The three image attributes can each be represented by an indicator vector, so the covariates (attributes) have a three-way tensor-variate structure. Our objective is to train a linear model to help us distinguish the visual characteristics of different attributes. Transforming the 3D tensors into vectors and fitting (1) requires a B of $250 \times 250 \times 3 \times 2 \times 3 \times 4$ or 4.5 million unconstrained parameters and an error covariance matrix Σ of over 17 billion similar parameters, making accurate inference (from only 605 observed images) impractical. Methodology that incorporates the reductions afforded by the tensor-variate structures of the responses and the covariates can redeem the situation. We revisit this dataset in Section 4.2.

1.2 Related Work and Overview of Our Contributions

The previous examples show the inadequacy of training (1) on tensor-valued data without additional accommodation for structure, as the sizes of B and Σ in unconstrained vector-variate regression grow with the dimensions of the tensor-valued responses and explanatory variables. Several regression frameworks that efficiently allow for tensor responses or covariates (but not both) have recently been considered [11], [12], [13], [14], [15], [16], [17]. Tensor-on-tensor regression (ToTR) refers to the case where both the response and covariates are tensors. In this context, [18] proposed an outer matrix product (OP) factorization of B, [19] suggested *canonical polyadic* or CANDECOMP/PARAFAC (CP) decomposition [20], [21], while [22] and [23] factorized B using a tensor ring (TR) [24], [25] format. The CP, TR and OP formats on B allow for quantum dimension reduction without affecting prediction or discrimination ability of the regression model. However, these methodologies do not account for dependence within tensor observations, the sampling distribution of their estimated coefficients and the natural connection that exists between ToTR and the related analysis of variance (ANOVA). Here, we propose a general ToTR framework that renders four low-rank tensor formats on the coefficient B: CP, Tucker (TK) [26], [27], [28], TR and the OP, while simultaneously allowing the errors to follow a *tensor-variate normal (TVN) distribution* [29], [30], [31], [32] that posits a Kronecker structure on the Σ of (1). Assuming TVN-distributed errors allows us to obtain the sampling distributions of the estimated coefficients under their assumed low-rank format. Indeed, Section 4.1 uses our derived sampling distributions to produce statistical parametric maps to help detect significant neurological interactions between death-related words and suicide attempter/ideator status. The TVN assumption on the errors also allows us to consider dependence within the tensor-valued observations. The Kronecker structured Σ in the TVN model renders a different covariance matrix for each tensor dimension, allowing us to simultaneously study multiple dependence contexts within the same framework. Here we also introduce the notion of tensor-variate ANOVA (TANOVA) under the ToTR framework, which is analogous to ANOVA and multivariate ANOVA (MANOVA) being instances of multiple linear regression (MLR) and MVMLR.

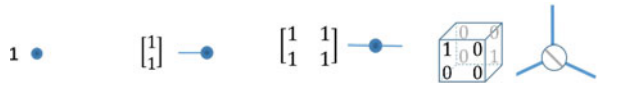
The rest of the paper is structured as follows. Section 2 first presents notations and network diagrams, low-rank tensor formats, the TVN distribution and our preliminary results that we develop for use in this paper. We formulate ToTR and TANOVA methodology with low-rank tensor formats on the covariates, and TVN errors. We provide algorithms for finding maximum likelihood (ML) estimators and study their properties. Section 3 evaluates performance of our methods in two simulation scenarios while Section 4 applies our methodology to the motivating applications of Sections 1.1.1 and 1.1.2. We conclude with some discussion. An online supplement with sections, theorems, lemmas, figures and equations prefixed with "S" is also available.

2 Theory and Methods

This section introduces a regression model with TVN errors and tensor-valued responses and covariates. We provide notations and definitions, then introduce our models and develop algorithms for ML estimation under the TK, CP, OP and TR low-rank formats. A special case leads us to TANOVA. We also derive asymptotic properties of our estimators and computational complexity of our algorithms.

2.1 Background and Preliminary Results

We provide a unified treatment of tensor reshapings and contractions by integrating the work of [27], [33], and [34] with our own results that we use later. We use $\mathrm{tr}(.)$, $(.)'$, and $(.)^-$ to denote the trace, transpose, and pseudo-inverse, I_n for the $n \times n$ identity matrix and $\otimes$ for the Kronecker product (Section S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3164836, has additional definitions and illustrations). We define tensors as multi-dimensional arrays of numbers. The total number of *modes* or sides of a tensor is called its *order*. We use lower-case letters (i.e., x) to specify scalars, bold lower-case italics (i.e., $\boldsymbol{x}$) for vectors, upper-case italics (i.e., X) for matrices, and calligraphic scripts (i.e., $\mathcal{X}$) for higher-order tensors. Random matrices or vectors are denoted using $\boldsymbol{X}$ and random tensors by $\boldsymbol{\mathcal{X}}$. We denote the $(i_1, \ldots, i_p)th$ element of a pth order tensor $\mathcal{X}$ using $\mathcal{X}(i_1, \ldots, i_p)$ or $\mathcal{X}(\boldsymbol{i})$ where $\boldsymbol{i} = [i_1, \ldots, i_p]'$. The vector



(a) $x \in \mathbb{R}$ (b) $\boldsymbol{x} \in \mathbb{R}^2$ (c) $X \in \mathbb{R}^{2\times 2}$ (d) $\mathcal{X} \in \mathbb{R}^{2\times 2\times 2}$

Fig. 1. Tensor network diagrams of a (a) scalar, (b) vector, (c) matrix and (d) third-order diagonal tensor.

outer product, with notation $\circ$, of p vectors generates the pth-order tensor $\mathcal{X} = \circ_{j=1}^{p} \boldsymbol{x}_j$ with $(i_1, \ldots, i_p)$th element $\mathcal{X}(\boldsymbol{i}) = \prod_{j=1}^{p} \boldsymbol{x}_j(i_j)$. Any pth-order tensor $\mathcal{X} \in \mathbb{R}^{m_1 \times \ldots \times m_p}$ (or $\mathbb{R}^{\times_{j=1}^{p} m_j}$) can be expressed using the vector outer product as

$$\mathcal{X} = \sum_{i_1=1}^{m_1} \cdots \sum_{i_p=1}^{m_p} \mathcal{X}(i_1, \ldots, i_p) \left(\bigcirc_{q=1}^{p} \boldsymbol{e}_{i_q}^{m_q} \right), \tag{2}$$

where $\boldsymbol{e}_i^m \in \mathbb{R}^m$ is a unit basis vector with 1 as the ith element and 0 everywhere else. Equation (2) allows us to reshape a tensor by only manipulating the vector outer product. A pth-order diagonal tensor $\mathbb{I}_r^p \in \mathbb{R}^{\times_{j=1}^{p} r}$ has ones where the indices in each mode coincide, and zeroes elsewhere, that is,

$$\mathbb{I}_r^p = \sum_{i=1}^{r} \left(\bigcirc_{q=1}^{p} \boldsymbol{e}_i^r \right). \tag{3}$$

Tensor structures are conveniently represented by tensor network diagrams that are a recent adaptation [34] from quantum physics where they were originally introduced to visually describe many-body problems. Each node in a tensor network diagram corresponds to a tensor and each edge coming from the node represents a mode. A node with no edges is a scalar, a node with one edge is a vector and a node with two edges is a matrix. More generally, a node with p edges is a pth-order tensor (Figs. 1a, 1b, 1c, and 1d). (The angle between edges has no meaning beyond aesthetics.) Diagonal tensors as in (3) are represented by putting a diagonal across the node, as in Fig. 1d.

2.1.1 Tensor Reshapings, Contractions, Low-Rank Formats

The matricization of a tensor is a matrix with its elements arranged differently. The following definition is from [33]

Definition 2.1. *Let $\mathcal{S} = \{r_1, \ldots, r_L\}$ and $\mathcal{T} = \{m_1, \ldots, m_M\}$ be ordered sets that partition the set of modes $\mathcal{M} = \{1, \ldots, p\}$ of $\mathcal{X} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$. Here $L + M = p$. Then if $\boldsymbol{i} = [i_1, \ldots, i_p]'$, the matricization $\mathcal{X}_{(\mathcal{S}\times\mathcal{T})}$ is a matrix of size $(\prod_{q\in\mathcal{S}} m_q) \times (\prod_{q\in\mathcal{T}} m_q)$ defined as*

$$\mathcal{X}_{(\mathcal{S}\times\mathcal{T})} = \sum_{i_1=1}^{m_1} \cdots \sum_{i_p=1}^{m_p} \mathcal{X}(\boldsymbol{i}) \left(\bigotimes_{q\in\mathcal{S}} \boldsymbol{e}_{i_q}^{m_q} \right) \left(\bigotimes_{q\in\mathcal{T}} \boldsymbol{e}_{i_q}^{m_q} \right)'. \tag{4}$$

We define matricizations in *reverse lexicographic* order to be consistent with the traditional matrix vectorization. This means that the q modes in the multiple Kronecker product (4) are selected in reverse order. Table 1 defines several reshapings by selecting different partitions $(\mathcal{S}, \mathcal{T})$ of $\mathcal{M}$. These definitions are clarified in (S3), (S4) and (S5), available online.

TABLE 1
Tensor Reshapings Defined by Specifying Partitions $(\mathcal{S}, \mathcal{T})$ of $\mathcal{M}$ in (4)

Reshaping	Notation	$\mathcal{S}$	$\mathcal{T}$
kth mode matricization	$\mathcal{X}_{(k)}$	$\{k\}$	$\{1, \ldots, k-1, k+1, \ldots, p\}$
kth canonical matricization	$\mathcal{X}_{<k>}$	$\{1, \ldots, k\}$	$\{k+1, \ldots, p\}$
vectorization	$\text{vec}(\mathcal{X})$	$\{1, \ldots, p\}$	$\varnothing$

Tensor contractions [34] generalize the matrix product to higher-ordered tensors. We use $\mathcal{X} \times_{k_1,\ldots,k_a}^{l_1,\ldots,l_a} \mathcal{Y}$ to denote the mode-$\binom{l_1,\ldots,l_a}{k_1,\ldots,k_a}$ product or contraction between the $(k_1, \ldots, k_a)$ modes of $\mathcal{X} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ and the $(l_1, \ldots, l_a)$ modes of $\mathcal{Y} \in \mathbb{R}^{\times_{j=1}^{q} n_j}$, where $m_{k_1} = n_{l_1}, \ldots, m_{k_a} = n_{l_a}$. This contraction results in a tensor of order $p + q - 2a$ where the a pairs of modes (l_j, k_j) get *contracted*. A simple contraction between the kth mode of $\mathcal{X}$ and the lth mode of $\mathcal{Y}$ has $(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_p, j_1, \ldots, j_{l-1}, j_{l+1}, \ldots, j_q)$th element

$$\sum_{t=1}^{m_l} \mathcal{X}(i_1, \ldots, i_{k-1}, t, i_{k+1}, \ldots, i_p) \mathcal{Y}(j_1, \ldots, j_{l-1}, t, j_{l+1}, \ldots, j_q). \tag{5}$$

Similarly, multiple contractions sum over multiple products of the elements of $\mathcal{X}$ and $\mathcal{Y}$. Table 2 defines some contractions using this notation. An important special case is *partial contraction* that contracts all the $p < q$ modes of $\mathcal{X} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ with the first p modes of $\mathcal{Y} \in \mathbb{R}^{\times_{j=1}^{q} m_j}$ resulting in a tensor $\langle \mathcal{X} | \mathcal{Y} \rangle = \mathcal{X} \times_{1,\ldots,p}^{1,\ldots,p} \mathcal{Y}$ of size $\mathbb{R}^{\times_{j=p+1}^{q} m_j}$. The partial contraction helps define ToTR, and can also be written as a matrix-vector multiplication using Lemma 2.1 (e) (below).

The tensor trace is a self-contraction between the two outer-most modes of a tensor. If $m_1 = m_p$, then

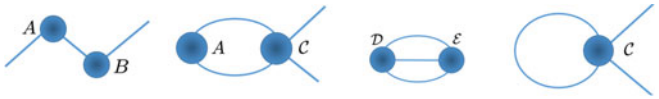
$$\text{tr}(\mathcal{X}) = \sum_{i=1}^{m_1} \mathcal{X}(i, :, : \ldots, :, i), \tag{6}$$

whence $\text{tr}(\mathcal{X}) \in \mathbb{R}^{\times_{j=2}^{p+1} m_j}$. The contraction between two distinct modes (from possibly the same tensor) is represented in tensor network diagrams by joining the corresponding edges (see Fig. 2 for examples). Also, applying the kth mode matricization to every mode of $\mathcal{V} \in \mathbb{R}^{\times_{j=1}^{p} g_j}$ with respect to (WRT) $A_i \in \mathbb{R}^{m_i \times g_i}, i = 1, 2, \ldots, p$ results in the TK product $\mathcal{B} = [\![\mathcal{V}; A_1, \ldots, A_p]\!]$ defined as

TABLE 2
Tensor Contractions, Where the Contraction Along One Mode is Defined as Per (5)

Contraction	Notation	Definition	Conditions
matrix product	XY	$X \times_2^1 Y$	$p = q = 2$
kth mode matrix product	$\mathcal{X} \times_k Y$	$\mathcal{X} \times_k^2 Y$	$q = 2$
kth mode vector product	$\mathcal{X} \bar{\times}_k \boldsymbol{y}$	$\mathcal{X} \times_k^1 \boldsymbol{y}$	$q = 1$
inner product	$\langle \mathcal{X}, \mathcal{Y} \rangle$	$\mathcal{X} \times_{1,\ldots,p}^{1,\ldots,q} \mathcal{Y}$	$p = q$
partial contraction	$\langle \mathcal{X} \vert \mathcal{Y} \rangle$	$\mathcal{X} \times_{1,\ldots,p}^{1,\ldots,p} \mathcal{Y}$	$p < q$
last mode with first mode	$\mathcal{X} \times^1 \mathcal{Y}$	$\mathcal{X} \times_p^1 \mathcal{Y}$	—

Here $\mathcal{X} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$, $\mathcal{Y} \in \mathbb{R}^{\times_{j=1}^{q} n_j}$, and X and Y are the cases where $p = 2$ and $q = 2$ respectively.



(a) $AB \in \mathbb{R}^{p\times r}$ (b) $\langle A|\mathcal{C}\rangle \in \mathbb{R}^{r\times p}$ (c) $\langle \mathcal{D},\mathcal{E}\rangle \in \mathbb{R}$ (d) $\mathrm{tr}(\mathcal{C}) \in \mathbb{R}^{q\times r}$

Fig. 2. Tensor network diagrams of (a) matrix product, (b) partial contraction, (c) inner product and (d) trace. Here $A \in \mathbb{R}^{p\times q}$, $B \in \mathbb{R}^{q\times r}$, $\mathcal{D},\mathcal{E} \in \mathbb{R}^{p\times q\times r\times p}$ and $\mathcal{C} \in \mathbb{R}^{p\times q\times r\times p}$.

$$\mathcal{B} = \sum_{i_1=1}^{m_1} \dots \sum_{i_p=1}^{m_p} \mathcal{V}(i_1 \dots i_p)\Big(\circ_{q=1}^{p} A_q(:, i_q)\Big). \tag{7}$$

A tensor $\mathcal{B}$ that can be written as the product in (7) is said to have a *TK format* with *TK rank*$(g_1, \dots, g_p)$. In this case, $\mathcal{V}$ is called the *core tensor* and $A_1, \dots, A_p$ are the *factor matrices*. When $g_i << m_i$ for $i = 1, \dots, p$, the TK format substantially reduces the number of unconstrained elements in a tensor and its complexity. TK formats are often fit by higher-order singular value (HOSVD) [35] or LQ (HOLQ) decomposition [36]. For a diagonal core tensor $\mathcal{V}$, as in (3), the TK format reduces to the CP format of rank r. This reduction is equivalent to setting the tensor $\mathcal{B}$ as the sum of r vector outer products, where the vectors correspond to the columns of matrix factors $A_i \in \mathbb{R}^{m_i\times r}, i = 1, \dots, p$,

$$\mathcal{B} = [\![\lambda; A_1, \dots, A_p]\!] = \sum_{i=1}^{r} \lambda(i)\Big(\overset{p}{\underset{q=1}{\circ}} A_q(:, i)\Big). \tag{8}$$

The vector $\lambda \in \mathbb{R}^r$ contains the diagonal entries of the core tensor, and is often set to the proportionality constants that make the matrix factors have unit column norms. When λ is ignored in the specification of (8), we assume that $\lambda = [1, 1, \dots, 1]'$. A tensor $\mathcal{B}$ is said to have an OP format if it can be written as $\mathcal{B} = \circ[\![A_1, A_2 \dots, A_p]\!]$, or

$$\mathcal{B} = \sum_{\substack{i_1,\dots,i_p \\ j_1\dots,j_p}} \Big(\prod_{q=1}^{p} A_q[i_q, j_q]\Big)\Big\{\Big(\overset{p}{\underset{q=1}{\circ}} \boldsymbol{e}_{j_q}^{h_q}\Big) \circ \Big(\overset{p}{\underset{q=1}{\circ}} \boldsymbol{e}_{i_q}^{m_q}\Big)\Big\}, \tag{9}$$

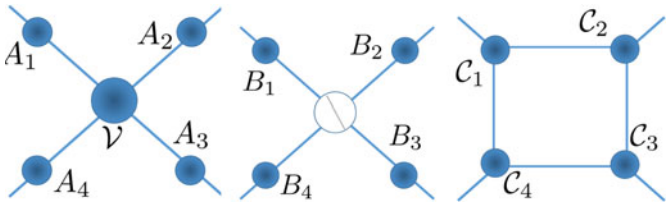
where for all $q = 1, \dots, p$, we have $A_q \in \mathbb{R}^{m_q\times h_q}$ and the summation over i_q is from 1 through m_q, and that over j_q is from 1 through h_q. Our novel OP format is essentially the outer product of multiple matrices, and is useful for expressing the TK product of (7) as a partial contraction between $\mathcal{V}$ and $\circ[\![A_1, A_2 \dots, A_p]\!]$, as we shortly state and prove in Theorem 2.1(b). The derivation needs some properties of tensor products and reshapings that we prove first in Lemma 2.1, along with several other properties that are useful for tensor manipulations. (Lemma 2.1(a),(c) and (d) have been stated without proof in [27], [33], and [34] but we provide proofs here for completeness.)

Lemma 2.1. *Let $\mathcal{X} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$. Then using the notation of Tables 1 and 2, where $k = 1, \dots, p$,*

(a) $\mathcal{X}_{<p-1>} = \mathcal{X}'_{(p)}$.
(b) $\mathrm{vec}(\mathcal{X}) = \mathrm{vec}(\mathcal{X}_{(1)}) = \mathrm{vec}(\mathcal{X}_{<1>}) = \dots = \mathrm{vec}(\mathcal{X}_{<p>})$.
(c) $\langle \mathcal{X}, \mathcal{Y}\rangle = (\mathrm{vec}\mathcal{X})'(\mathrm{vec}\mathcal{Y}) = \mathrm{tr}(\mathcal{X}_{(k)}\mathcal{Y}'_{(k)})$, $\mathcal{Y} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$.
(d) $\mathrm{vec}[\![\mathcal{X}; A_1, \dots, A_p]\!] = (\bigotimes_{i=p}^{1} A_i)\mathrm{vec}(\mathcal{X})$, *where* $A_i \in R^{n_i\times m_i}$ *for any* $n_i \in \mathbb{N}$.
(e) $\mathrm{vec}\langle \mathcal{X}|\mathcal{B}\rangle = \mathcal{B}'_{<p>}\mathrm{vec}(\mathcal{X}), \mathcal{B} \in \mathbb{R}^{(\times_{j=1}^{p} m_j)\times(\times_{j=1}^{q} h_j)}$.
(f) $\mathrm{vec}(\mathcal{X}_{(k)}) = K_{(k)}\mathrm{vec}(\mathcal{X})$, *where* $K_{(k)} = (I_{\prod_{i=k+1}^{p} m_i} \otimes K_{\prod_{i=1}^{k-1} m_i, m_k})$.

Proof. See Section S1.3, available online. □

We now use Lemma 2.1 to state and prove the following



(a) $[\![\mathcal{V}; A_1, A_2, A_3, A_4]\!]$ (b) $[\![B_1, B_2, B_3, B_4]\!]$ (c) $\mathrm{tr}(\mathcal{C}_1\times^1\mathcal{C}_2\times^1\mathcal{C}_3\times^1\mathcal{C}_4)$

Fig. 3. Tensor network diagrams of example fourth-order tensor of (a) TK format (7), (b) CP format (8), and (c) TR format (11).

Theorem 2.1. *Consider a pth-order tensor $\mathcal{X}$ and matrices $M_1, \dots, M_p$ such that the TK product with $\mathcal{X}$ can be formed. Then*

(a) $\circ[\![M_1, \dots, M_p]\!]_{<p>} = \bigotimes_{q=p}^{1} M'_q$.
(b) $\langle \mathcal{X}| \circ [\![M_1, \dots, M_p]\!]\rangle = [\![\mathcal{X}; M_1, \dots, M_p]\!]$.
(c) For any $k = 1, \dots p$, let $\mathcal{S} = \{k, p + k\}$. Then

$$\begin{aligned} &\circ [\![M_1, \dots, M_p]\!]_{(\mathcal{S}\times\mathcal{S}^c)} \\ &= (\mathrm{vec}M'_k)(\mathrm{vec} \circ [\![M_1, \dots, M_{k-1}, M_{k+1}, \dots, M_p]\!])'. \end{aligned} \tag{10}$$

Proof. See Section S1.4, available online. □

Remark: If $\mathcal{B} = \circ[\![X', X']\!]$ for some matrix X, then Theorem 2.1(a) implies that $\mathcal{B}_{<2>} = X \otimes X$ while $\mathcal{B}_{(1,3)\times(2,4)} = (\mathrm{vec}X)(\mathrm{vec}X)'$ by Theorem 2.1(c). In other words, $X \otimes X$ and $(\mathrm{vec}X)(\mathrm{vec}X)'$ are different matricizations of the same OP tensor $\mathcal{B}$, which formalizes our intuition because both $X \otimes X$ and $(\mathrm{vec}X)(\mathrm{vec}X)'$ have the same number of elements, and it motivates naming the format *outer product*.

Finally, a tensor $\mathcal{B}$ is said to have a Tensor Ring (TR) format with TR rank $(g_1, \dots, g_p)$ if it can be expressed as

$$\mathcal{B} = \mathrm{tr}(\mathcal{G}_1 \times^1 \dots \times^1 \mathcal{G}_p), \tag{11}$$

where $\mathcal{G}_i \in \mathbb{R}^{g_{i-1}\times m_i\times g_i}$ for $i = 1, \dots, p$ and $g_0 = g_p$. The TR format is called the Matrix Product State (MPS) with closed boundary conditions in many-body physics [37]. When exactly one of the TR ranks is 1, the TR format is the same as the Tensor Train (TT) format [25] and is the MPS with open boundary conditions.

We conclude our discussion of low-rank tensor formats by using tensor network diagrams to illustrate in Fig. 3, a fourth-order tensor in the TK, CP and TR formats.

2.1.2 The TVN Distribution

The matrix-variate normal distribution, abbreviated in [38] as MxVN to distinguish it from the vector-variate multinormal distribution (MVN), was studied extensively in [39]. A random matrix $\boldsymbol{X} \in \mathbb{R}^{m_1\times m_2}$ follows a MxVN distribution if $\mathrm{vec}(\boldsymbol{X})$ is MVN with covariance matrix $\Sigma_2 \otimes \Sigma_1$, where $\Sigma_k \in \mathbb{R}^{m_k\times m_k}$ for $k = 1, 2$. The TVN distribution, formulated in Definition 2.2, extends this idea to the case of higher-order random tensors. For simplicity, we define the following notation for use in the rest of the paper:

$$m = \prod_{i=1}^{p} m_i,\ m_{-k} = m/m_k,\ \Sigma = \bigotimes_{i=p}^{1} \Sigma_i,\ \Sigma_{-k} = \bigotimes_{\substack{i=p \\ i\neq k}}^{1} \Sigma_i.$$

Definition 2.2. *A random tensor* $\mathcal{X} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ *follows a pth-order TVN distribution with mean* $\mathcal{M} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ *and non-negative definite scale matrices* $\Sigma_i \in \mathbb{R}^{m_i \times m_i}$ *for* $i = 1, 2, \ldots, p$ *(i.e.,* $\mathcal{X} \sim \mathcal{N}_{\boldsymbol{m}}(\mathcal{M}, \Sigma_1, \Sigma_2 \ldots, \Sigma_p)$ *where* $\boldsymbol{m} = [m_1, m_2, \ldots, m_p]'$ *) if* $\text{vec}(\mathcal{X}) \sim \mathcal{N}_m(\text{vec}(\mathcal{M}), \Sigma)$.

The Kronecker product in Definition 2.2 is in reverse order because we have defined vectorization in reverse lexicographic order. Definition 2.2 defines the TVN distribution in terms of a vectorization. We state and prove the distribution of other tensor reshapings in Theorem 2.2. These results are essential in the development of ToTR models with TVN errors, as they allow us to model the vectorized tensor errors in terms of the MVN distribution.

Theorem 2.2. *The following statements are equivalent:*

(a) $\mathcal{Y} \sim \mathcal{N}_{\boldsymbol{m}}(\mathcal{M}, \Sigma_1, \Sigma_2, \ldots, \Sigma_p)$
(b) $\text{vec}(\mathcal{Y}) \sim \mathcal{N}_m(\text{vec}(\mathcal{M}), \Sigma)$
(c) $\mathcal{Y}_{(k)} \sim \mathcal{N}_{[m_k, m_{-k}]'}(\mathcal{M}_{(k)}, \Sigma_k, \Sigma_{-k}), \quad k = 1, 2, \ldots, p$
(d) For $k = 1, 2, \ldots, p$ *and* $\boldsymbol{m}_k = [\prod_{i=1}^{k} m_i, \prod_{i=k+1}^{p} m_i]'$,
$\mathcal{Y}_{<k>} \sim \mathcal{N}_{\boldsymbol{m}_k}(\mathcal{M}_{<k>}, \bigotimes_{i=k}^{1} \Sigma_i, \bigotimes_{i=p}^{k+1} \Sigma_i)$

Proof. (a) and (b) are equivalent, following Definition 2.2 while (b) and (c) are so from Lemma 2.1(f) with $K_{(k)} \Sigma K'_{(k)} = \Sigma_{-k} \otimes \Sigma_k$. Further, (b) and (d) are equivalent because of Lemma 2.1(b). □

The density of $\mathcal{Y} \sim \mathcal{N}_{\boldsymbol{m}}(\mathcal{M}, \Sigma_1, \Sigma_2, \ldots, \Sigma_p)$ is $f(\mathcal{Y}; \mathcal{M}, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2} D^2_{\Sigma}(\mathcal{Y}, \mathcal{M})\}$, where $D^2_{\Sigma}(\mathcal{Y}, \mathcal{M})$ is the squared Mahalanobis distance between $\mathcal{Y}$ and $\mathcal{M}$, and has the equivalent representations

$$D^2_{\Sigma}(\mathcal{Y}, \mathcal{M}) = \text{vec}(\mathcal{Y} - \mathcal{M})' \Sigma^{-1} \text{vec}(\mathcal{Y} - \mathcal{M}) \equiv \langle (\mathcal{Y} - \mathcal{M}), [\![(\mathcal{Y} - \mathcal{M}); \Sigma_1^{-1}, \Sigma_2^{-1}, \ldots, \Sigma_p^{-1}]\!] \rangle, \quad (12)$$

by Lemmas 2.1(c) and 2.1(d). Property S1.2, available online, provides similar alternative expressions for the determinant $\det(\Sigma)$ of Σ.

2.2 Tensor-Variate Linear Models With TVN Errors

2.2.1 Tensor-on-Tensor Regression

We formulate the ToTR model as

$$\mathcal{Y}_i = \Upsilon + \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad i = 1, 2 \ldots, n, \quad (13)$$

where the response $\mathcal{Y}_i \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ and the covariate $\mathcal{X}_i \in \mathbb{R}^{\times_{j=1}^{l} h_j}$ are both tensor-valued, $\mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{\boldsymbol{m}}(0, \sigma^2\Sigma_1, \Sigma_2, \ldots, \Sigma_p)$ is the TVN-distributed error, $\mathcal{B} \in \mathbb{R}^{(\times_{j=1}^{l} h_j) \times (\times_{j=1}^{p} m_j)}$ is the (tensor-valued) regression parameter and Υ is the (tensor-valued) intercept. This model is essentially the classical MVMLR model but exploits the tensor-variate structure of the covariates and responses to reduce the total number of parameters. To see this, we apply Lemma 2.1(e) and Theorem 2.2(b) to vectorize (13) as $\text{vec}(\mathcal{Y}_i) = \text{vec}(\Upsilon) + \mathcal{B}'_{<l>} \text{vec}(\mathcal{X}_i) + \boldsymbol{e}_i$, where $\boldsymbol{e}_i \overset{iid}{\sim} \mathcal{N}_m(0, \sigma^2\Sigma)$ is the error. This formulation leads to a MVMLR model with intercept, which can be incorporated into the covariates as $[\text{vec}(\Upsilon)\mathcal{B}'_{<l>}][1, \text{vec}(\mathcal{X}_i)']'$. But the covariate $[1, \text{vec}(\mathcal{X}_i)']'$ is then no longer a vectorized tensor and we can not exploit the tensor structure of $\mathcal{X}_i$. To obviate this possibility, (13) includes a separate intercept term Υ.

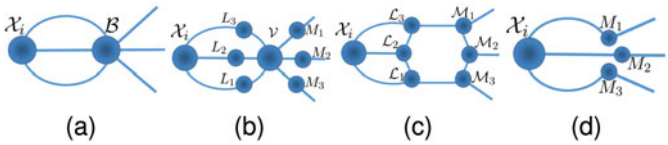


Fig. 4. Tensor network diagrams of tensor-on-tensor regression when both the response and the covariates are third-order tensors and for (a) $\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle$ and (b-d) the special cases when the $\mathcal{B}$ in $\langle \mathcal{X}_i | \mathcal{B} \rangle$ are of (b) Tucker, (c) TR and (d) OP formats (illustrated in Fig. 3).

Imposing a low-rank format on $\mathcal{B}$ proffers several advantages. First, it makes the regression model practical to use, as accurate estimation of an unstructured version of $\mathcal{B}$ may otherwise be prohibitive when dimensionality is high relative to sample size. Second, $\mathcal{B}$ can be interpreted, in spite of its high dimensions, as being explainable through a few lower-dimensional tensor factor matrices (Fig. 4). These explanations mirror the many-body problem in physics, where weakly-coupled degrees of freedom are often embedded in ultra-high-dimensional Hilbert spaces [37], [40]. We now turn to the problem of learning $\mathcal{B}$ in the given setup.

The dispersion matrix Σ in the TVN distribution specification of the $\mathcal{E}_i$s in (13) is Kronecker-separable and leads to the number of unconstrained parameters from $(\prod_{i=1}^{p} m_i) \times (\prod_{i=1}^{p} m_i + 1)/2$ to $\sum_{i=1}^{p} m_i(m_i + 1)/2$. Further, this Kronecker-separable structure is intuitive because it assigns a covariance matrix to each tensor-response dimension. This allows us to separately incorporate different dependence contexts that exist within a tensor. For example, a tensor may have temporal and spatial contexts in its modes. Kronecker separability allows us to assign a covariance matrix to each of these different contexts. However, Kronecker separability results in unidentifiable scale matrices (Σ_i), as $cA \otimes B = A \otimes cB$ for any matrices A, B and $c \neq 0$, so we constrain the scale matrices to each have $\Sigma_i(1,1) = 1$, and introduce a parameter σ^2 to capture an overall proportional scalar variance. This approach further reduces the number of parameters by $p - 1$ and imposes a curved exponential family distribution on the errors [41].

2.2.2 The TANOVA Model

Suppose that we observe independent tensors $\mathcal{Y}_{j_1, \ldots, j_l, i} \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ $(i = 1, \ldots, n_{j_1, \ldots, j_l})$, where $j_k (k = 1, \ldots, l)$, indexes the kth categorical (factor) variable of h_k levels, that is, $j_k \in \{1, \ldots, h_k\}$. The tensor-valued parameter $\mathcal{B}$ encoding the dimensions of $\mathcal{Y}_{i, j_1, \ldots, j_l}$ and all the possible factor classes is of size $h_1 \times \ldots \times h_l \times m_1 \times \ldots \times m_p$. To see this, consider the lth ordered single-entry tensor $\mathcal{X}_{j_1, \ldots, j_l} = \circ_{q=1}^{l} \boldsymbol{e}_{i_q}^{h_q}$ that is unity at $(j_1, \ldots, j_l)$ and zero everywhere else. Then $\langle \mathcal{X}_{j_1, \ldots, j_l} | \mathcal{B} \rangle \in \mathbb{R}^{\times_{j=1}^{p} m_j}$ and

$$\langle \mathcal{X}_{j_1, \ldots, j_l} | \mathcal{B} \rangle = \mathcal{B}[j_1, \ldots, j_l, :, :, \ldots, :]. \quad (14)$$

Therefore, modeling $\mathbb{E}(\mathcal{Y}_{i, j_1, \ldots, j_l})$ as $\langle \mathcal{X}_{j_1, \ldots, j_l} | \mathcal{B} \rangle$ results in each factor combination $(j_1, \ldots, j_p)$ getting assigned its own mean parameter as a sub-tensor of $\mathcal{B}$. This high-dimensional parameter $\mathcal{B}$ is identical to cell-means MANOVA if we vectorize (14) using Lemma 2.1(e) as $\mathcal{B}'_{<l>} \text{vec}(\mathcal{X}_{j_1, \ldots, j_l}) =$

$\mathrm{vec}(\mathcal{B}[j_1,\ldots,j_p,:,:,\ldots,:])$, in which case $\mathrm{vec}(\mathcal{X}_{j_1,\ldots,j_p}) = \otimes_{q=l}^{1} \boldsymbol{e}_{i_q}^{h_q}$ corresponds to a row in the MANOVA design matrix. Although this formulation is fundamentally the same as (14), the latter helps us visualize and formulate a low-rank format on $\mathcal{B}$. Based on the format, we refer to tensor-valued response regression models with the mean in (14) as TANOVA(l,p), where l is the number of different factors and p is the order of the tensor-valued response. In this way, because scalar and vector variables are tensors of order 0 and 1, ANOVA and MANOVA correspond to TANOVA(1,0) and TANOVA(1,1), respectively. In general, TANOVA(l,p) with TVN errors can be expressed as

$$\mathcal{Y}_i = \langle \mathcal{X}_i | \mathcal{B} \rangle + \mathcal{E}_i, \quad \mathcal{E}_i \overset{iid}{\sim} \mathcal{N}_{\boldsymbol{m}}(0, \sigma^2\Sigma_1, \Sigma_2, \ldots, \Sigma_p), \tag{15}$$

where $\mathcal{X}_i$ is the single-entry tensor that contains all the assigned factors of $\mathcal{Y}_i$, for $i = 1,\ldots,n$. Model (15) is a ToTR model as in (13) with no intercept. This is analogous to ANOVA and MANOVA being special cases of univariate and multivariate multiple linear regressions, respectively. Further, the log-likelihood function of (15) is

$$\ell = -\frac{n}{2}\log|2\pi\sigma^2\Sigma| - \frac{1}{2\sigma^2}\sum_{i=1}^{n} D_{\Sigma}^2(\mathcal{Y}_i, \langle \mathcal{X}_i|\mathcal{B}\rangle). \tag{16}$$

2.3 Parameter Estimation

We obtain estimators of (13) before deriving their properties.

2.3.1 Profiling the Intercept

We first show that the intercept in (13) can be profiled out by centering the covariates and the responses. To see this, we express the loglikelihood in terms of Υ as

$$\ell = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} D_{\Sigma}^2(\mathcal{Y}_i, \Upsilon + \langle \mathcal{X}_i|\mathcal{B}\rangle). \tag{17}$$

We define the tensor differential of an inner product WRT Υ using the matrix differential $\langle \partial\Upsilon, \mathcal{S}\rangle = \mathrm{tr}(\partial\Upsilon_{(1)}\mathcal{S}'_{(1)})$. Applying it to (17) yields

$$\partial\ell(\Upsilon) = \frac{1}{\sigma^2}\langle \partial\Upsilon, [\![\mathcal{S}; \Sigma_1^{-1},\ldots,\Sigma_p^{-1}]\!]\rangle,$$

where $n^{-1}\mathcal{S} = \bar{\mathcal{Y}} - \langle \bar{\mathcal{X}}|\mathcal{B}\rangle - \Upsilon$ and $\bar{\mathcal{Y}}, \bar{\mathcal{X}}$ are the averaged responses and covariates. Now, $\partial\ell(\Upsilon) = 0$ if $\mathcal{S} = 0$, and for fixed $\mathcal{B}$, the ML estimator (MLE) of Υ is $\widehat{\Upsilon}(\mathcal{B}) = \bar{\mathcal{Y}} - \langle \bar{\mathcal{X}}|\mathcal{B}\rangle$. Setting Υ in (17) to be $\widehat{\Upsilon}(\mathcal{B})$ yields (16) with centered responses and covariates, so we assume without loss of generality (WLOG) that (15) has no intercept, and estimate the other parameters. Our estimation uses block-relaxation [42] to optimize (16): we partition the parameter space into blocks and serially optimize the parameters in each block while holding fixed the other parameters.

2.3.2 Estimation of $\mathcal{B}, \Sigma_1, \Sigma_2, \ldots, \Sigma_p$ Given σ^2

Our estimates simplify as per the format of $\mathcal{B}$ so we consider each case individually, before providing an overview.

TK format. Let $\mathcal{B}$ have TK format of rank $(c_1,\ldots,c_l, d_1,\ldots,d_p)$

$$\mathcal{B}_{TK} = [\![\mathcal{V}; L_1,\ldots,L_l, M_1,\ldots,M_p]\!], \tag{18}$$

where $M_k'\Sigma_k^{-1}M_k = I_{d_k}$ for $k = 1,\ldots,p$. Then the number of parameters to be estimated goes down from the unconstrained $\prod_{i=1}^{l} h_i \prod_{i=1}^{p} m_i$ to $\prod_{i=1}^{l} c_i \prod_{i=1}^{p} d_i + \sum_{i=1}^{l} c_i h_i + \sum_{i=1}^{p} d_i m_i$. The constraint $M_k'\Sigma_k^{-1}M_k = I_{d_k}$ greatly simplifies estimation and inference. Using (18), we vectorize (15) for any $k = 1,\ldots,l$ as

$$\mathrm{vec}(\mathcal{Y}_i) = \mathcal{H}^{TK}_{ik<2>}\mathrm{vec}(L_k) + \boldsymbol{e}_i, \tag{19}$$

where $\mathcal{H}^{TK}_{ik<2>}$ is the 2-canonical matricization of the tensor

$$\mathcal{H}^{TK}_{ik} = \mathcal{X}_i \times^{1,\ldots,k-1,k+1,\ldots,l}_{1,\ldots,k-1,k+1,\ldots,l} [\![\mathcal{V}; L_1,\ldots,L_{k-1}, I_{h_k}, L_{k+1},\ldots,L_l, M_1,\ldots,M_p]\!], \tag{20}$$

and $\boldsymbol{e}_i$s are i.i.d $\mathcal{N}_m(0,\sigma^2\Sigma)$. Optimizing (16) WRT L_k forms its own block, which corresponds to a MVMLR model where, for $S_k^{TK} = \sum_{i=1}^{n}(\mathcal{H}^{TK}_{ik<2>}\Sigma^{-1}\mathcal{H}^{TK}_{ik<2>}{}')$,

$$\mathrm{vec}(\widehat{L_k}) = (S_k^{TK})^{-1}\left(\sum_{i=1}^{n}\mathcal{H}^{TK}_{ik<2>}\Sigma^{-1}\mathrm{vec}(\mathcal{Y}_i)\right). \tag{21}$$

The computation of S_k^{TK} is greatly simplified based on the constraint that $M_k'\Sigma_k^{-1}M_k = I_{d_k}$ for all $k = 1,2,\ldots,p$. For fixed $L_1,\ldots,L_p,\Sigma_1,\ldots,\Sigma_p$, we estimate $M_1,\ldots,M_p,\mathcal{V}$. We first show that $\mathcal{V}$ can be profiled from the loglikelihood for fixed $M_1,\ldots,M_p$. To see this, we write an alternative vectorized form of (15) as

$$\mathrm{vec}(\mathcal{Y}_i) = \left(\otimes_{i=p}^{1} M_i\right)\mathcal{V}'_{<l>}\boldsymbol{w}_i + \boldsymbol{e}_i,$$

where $\boldsymbol{w}_i = \mathrm{vec}[\![\mathcal{X}_i; L_1', L_2',\ldots,L_l']\!]$. Letting $Z = Y - M\mathcal{V}'_{<l>}W$, for $M = \otimes_{k=p}^{1} M_k$, $Y = [\mathrm{vec}\mathcal{Y}_1 \ldots \mathrm{vec}\mathcal{Y}_n]$, and $W = [\boldsymbol{w}_1 \ldots \boldsymbol{w}_n]$ simplifies (16) to

$$\ell = -\frac{n}{2}\log|\Sigma| - \frac{1}{2\sigma^2}\left\{\mathrm{tr}(Z'\Sigma^{-1}Z)\right\}. \tag{22}$$

Optimizing (22) for fixed $M_1,\ldots,M_p$ yields the profiled MLE

$$\widehat{\mathcal{V}}_{<l>}(M_1,\Sigma_1,\ldots,M_p,\Sigma_p) = W^{-'}Y'(\bigotimes_{k=p}^{1}\Sigma_k^{-1}M_k), \tag{23}$$

where W^- is the right inverse of W. Therefore, given values of all M_ks, we obtain $\widehat{\mathcal{V}}$ by simply inserting them in (23). To estimate M_k we profile $\widehat{\mathcal{V}}$ out of the loglikelihood by replacing (23) into (22), and expressing it up to a constant as

$$\ell(M_k,\Sigma_k) = \frac{1}{2\sigma^2}||M_k'\Sigma_k^{-1}Q_k||_2^2, \tag{24}$$

where $Q_k = [\![\mathcal{Y}_T; M_1'\Sigma_1^{-1},\ldots,M_{k-1}'\Sigma_{k-1}^{-1}, I_{m_k}, M_{k+1}'\Sigma_{k+1}^{-1},\ldots,M_p'\Sigma_p^{-1}, W^-W]\!]_{(k)}$ and $\mathcal{Y}_T \in \mathbb{R}^{(\times_{j=1}^{p} m_j)\times n}$ is such that $\mathcal{Y}_T(:,\ldots,:,i) = \mathcal{Y}_i$ for $i = 1,\ldots,n$. From (24), the MLE of M_k is obtained via generalized SVD of Q_k [43]

$$\widehat{M}_k(\Sigma_k) = \underset{M_k: M_k'\Sigma_k^{-1}M_k = I_{d_k}}{\arg\max} ||M_k'\Sigma_k^{-1}Q_k||_2^2 = \Sigma_k^{1/2}U, \tag{25}$$

with the leading d_k left singular vectors of $\Sigma_k^{-1/2}Q_k$ as the columns of U. To estimate Σ_k at fixed $\mathcal{B}_{TK}$, we write (22) as

$$\ell(\Sigma_k) = -\frac{nm_{-k}}{2}\log|\Sigma_k| - \frac{1}{2\sigma^2}\mathrm{tr}(\Sigma_k^{-1}S_k), \tag{26}$$

where $S_k = \sum_{i=1}^n \mathcal{Z}_{i(k)}\Sigma_{-k}^{-1}\mathcal{Z}_{i(k)}'$ and $\mathcal{Z}_{i(k)} = \mathcal{Y}_i - \langle \mathcal{X}_i|\mathcal{B}\rangle$. The MLE of Σ_k under the (TK format) constraint of (18) is

$$\widehat{\Sigma}_k = \underset{\Sigma_k:\Sigma_k[1,1]=1}{\arg\max}\ \ell(\Sigma_k) = ADJUST(nm_{-k}, \sigma^2, S_k), \tag{27}$$

where the *ADJUST* procedure is as introduced in [44] and that was shown to satisfy the Karush-Kuhn-TK (KKT) conditions. Without this constraint, the MLE is $S_k/(nm_{-k}\sigma^2)$. Further reductions can be obtained by imposing additional parameterized structures on Σ_ks, as needed. Our block relaxation algorithm, detailed in Algorithm 1, is initialized by methods discussed in Section 2.3.4.

Algorithm 1. Block-Relaxation Algorithm for ToTR (13) With TK Formatted $\mathcal{B}$.

Initial values $k=0, \widehat{\sigma}^{2(0)}, \widehat{L}_1^{(0)}, \ldots, \widehat{L}_l^{(0)}, (\widehat{M}_1^{(0)}, \widehat{\Sigma}_1^{(0)}), \ldots, (\widehat{M}_p^{(0)}, \widehat{\Sigma}_p^{(0)})$

Center the data while saving the means $\bar{\mathcal{X}}, \bar{\mathcal{Y}}$.
while *convergence criteria is not met* **do**
  for $j = 1, 2, \ldots, p$ **do**
    $\widehat{M}_j^{(k+1)}$ as per (25)
  end
  $\widehat{\mathcal{V}}^{(k+1)}$ as per (23)
  for $j = 1, 2, \ldots, l$ **do**
    $\widehat{L}_j^{(k+1)}$ as per (21)
  end
  for $j = 1, 2, \ldots, p$ **do**
    $\widehat{\Sigma}_j^{(k+1)}$ as per (27)
    $\widehat{\sigma}^{2(k+1)}$ as per (39)
  end
  $k \leftarrow k+1$
end
$\widehat{\Upsilon} = \bar{\mathcal{Y}} - \langle \bar{\mathcal{X}}|\widehat{\mathcal{B}}_{TK}\rangle$

CP format. We now optimize (16) when $\mathcal{B}$ is

$$\mathcal{B}_{CP} = [\![\lambda; L_1, L_2, \ldots, L_l, M_1, M_2, \ldots, M_p]\!], \tag{28}$$

that is, of CP format of rank r. Then, with $\Sigma_k = I_{m_k}$ for $k = 1, 2, \ldots, p$, (15) reduces to the framework of [19]. The CP format reduces the number of parameters in $\mathcal{B}$ from $\prod_{i=1}^p m_i \prod_{i=1}^l h_i$ to $r(\sum_{i=1}^p m_i + \sum_{i=1}^l h_i)$. Here also, we optimize (16) via a block-relaxation algorithm. The kth block corresponds to (M_k, Σ_k) for $k = 1, 2, \ldots, p$ and can be estimated in a MVMLR framework by applying Theorem 2.2(c) on the kth mode matricized form of (15)

$$\mathcal{Y}_{i(k)} = M_k G_{ik}^{CP} + E_i, \quad E_i \overset{iid}{\sim} \mathcal{N}_{m\times m_{-k}}(0, \sigma^2\Sigma_k, \Sigma_{-k}) \tag{29}$$

where $G_{ik}^{CP} \equiv \mathcal{G}_{ik(k)}^{CP}$ is the kth mode matricization of $\mathcal{G}_{ik}^{CP} = [\![\langle [\![\mathcal{X}_i; L_1', \ldots, L_l']\!] | \mathcal{I}_r^{p+l}\rangle; M_1, \ldots, M_{k-1}, I_{m_k}, M_{k+1}, \ldots, M_p]\!]$.

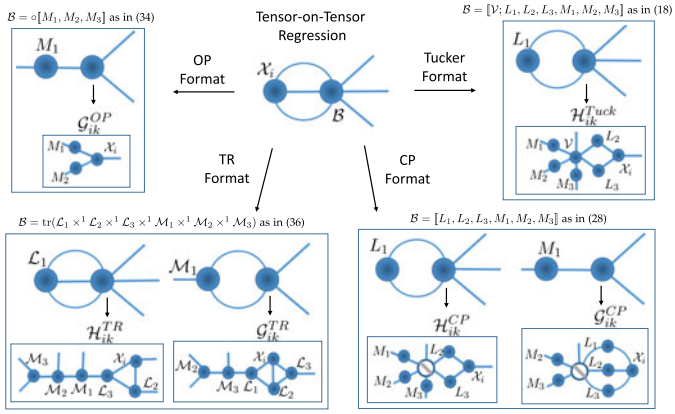


Fig. 5. Equivalent tensor-network diagrams for $\langle \mathcal{X}_i, \mathcal{B}\rangle$ when $p = l = 3$, which can be expressed in multiple ways depending on the tensor factor to be estimated and the type of low-rank on $\mathcal{B}$ (which are illustrated in Fig. 4). By choosing the tensors $\mathcal{G}_{ik}$ and $\mathcal{H}_{ik}$, the tensors M_1 and $\mathcal{M}_1$ can be estimated using multivariate multiple linear regression, and the tensors L_1 and $\mathcal{L}_1$ can be estimated using multiple linear regression, respectively. In these cases, matricized versions of $\mathcal{G}_{ik}$ and $\mathcal{H}_{ik}$ are part of the design matrix.

Additional simplifications of G_{ik}^{CP} are possible, for example, using (Section S1.5.1), available online, the Khatri-Rao product ($\odot$) [27]. When all parameters except (M_k, Σ_k) are held fixed, (29) matches a MVMLR model with loglikelihood

$$\ell(\Sigma_k, M_k) = \frac{nm_{-k}}{2}\log|\Sigma_k^{-1}| - \frac{1}{2\sigma^2}\mathrm{tr}(\Sigma_k^{-1}S_k), \tag{30}$$

with $S_k = \sum_{i=1}^n Z_{ik}\Sigma_{-k}^{-1}Z_{ik}'$, where $Z_{ik} = \mathcal{Y}_{i(k)} - M_k G_{ik}^{CP}$. Then the MLEs are

$$\widehat{M}_k = \sum_{i=1}^n \mathcal{Y}_{i(k)}\Sigma_{-k}^{-1}G_{ik}^{CP\prime}\left[\sum_{i=1}^n G_{ik}^{CP}\Sigma_{-k}^{-1}G_{ik}^{CP\prime}\right]^{-1},$$
$$\widehat{\Sigma}_k(M_k) = ADJUST(nm_{-k}, \sigma^2, S_k). \tag{31}$$

The matrices $\sum_{i=1}^n \mathcal{Y}_{i(k)}\Sigma_{-k}^{-1}G_{ik}^{CP\prime}$, $\sum_{i=1}^n G_{ik}^{CP}\Sigma_{-k}^{-1}G_{ik}^{CP\prime}$ are substantially simplified in Section S1.5.1. We estimate Σ_k by directly optimizing (30). The other l blocks in the block-relaxation algorithm correspond to $L_1, \ldots, L_l$ and are each MVMLR models obtained by vectorizing (15) as

$$\mathrm{vec}(\mathcal{Y}_i) = H_{ik}^{CP}\mathrm{vec}(L_k) + \boldsymbol{e}_i, \quad \boldsymbol{e}_i \overset{iid}{\sim} \mathcal{N}_m(0, \sigma^2\Sigma), \tag{32}$$

where $H_{ik}^{CP} = \mathcal{H}_{ik<2>}^{CP}$ and $\mathcal{H}_{ik}^{CP}$ is identical to the $\mathcal{H}_{ik}^{TK}$ of (20), but for the fact that $\mathcal{V}$ is the diagonal tensor $\mathcal{I}_r^{p+l}$. (Fig. 5 displays $\mathcal{H}_{ik}^{CP}$ for when $p=l=3$.) For $k = 1, \ldots, l$, holding all parameters except L_k fixed makes (32) a MLR model with the MLE of L_k obtained as

$$\mathrm{vec}(\widehat{L}_k) = \left(\sum_{i=1}^n H_{ik}^{CP}\Sigma^{-1}H_{ik}^{CP\prime}\right)^{-1}\left(\sum_{i=1}^n H_{ik}^{CP}\Sigma^{-1}\mathrm{vec}(\mathcal{Y}_i)\right). \tag{33}$$

The matrices $\sum_{i=1}^n H_{ik}^{CP}\Sigma^{-1}\mathrm{vec}(\mathcal{Y}_i)$, $\sum_{i=1}^n H_{ik}^{CP}\Sigma^{-1}H_{ik}^{CP\prime}$ are substantially simplified in Section S1.5.1, available online. As summarized in Fig. 5, the tensors $\mathcal{H}_{ik}^{CP}$ and $\mathcal{G}_{ik}^{CP}$ play a critical role in the estimation of M_k and L_k through (29) and (32), permitting the use of standard MVMLR and MLR estimation methods. From (8), we deduce that the jth columns of all the factor matrices in the CP decomposition (28) are

identifiable only up to a constant. So we constrain the columns to have unit norm. The MLEs of our parameters are obtained using a block-relaxation algorithm, as outlined in Algorithm 2.

Algorithm 2. Block-Relaxation Algorithm for ToTR (13) With CP Formatted $\mathcal{B}$.

Initial values $k=0, \widehat{\sigma}^{2(0)}, \widehat{L}_2^{(0)}, \widehat{L}_3^{(0)}, \ldots, \widehat{L}_l^{(0)}, \widehat{M}_1^{(0)}, \widehat{M}_2^{(0)}, \ldots, \widehat{M}_p^{(0)}$
1: Center the data while saving the means $\bar{\mathcal{X}}, \bar{\mathcal{Y}}$.
2: **while** *convergence criteria is not met* **do**
3: **for** $j=1,2,\ldots,l$ **do**
4: $\widehat{L}_j^{(k+1)}$ as per (33) and normalize its columns
5: **end**
6: **for** $j=1,2,\ldots,p-1$ **do**
7: $(\widehat{M}_k^{(k+1)}, \widehat{\Sigma}_k^{(k+1)})$ as per (31) and normalize the columns of $\widehat{M}_k^{(k+1)}$
8: **end**
9: $(\widehat{M}_p^{(k+1)}, \widehat{\Sigma}_p^{(k+1)})$ as per (31)
10: $\widehat{\sigma}^{2(k+1)}$ as per (39)
11: Normalize the columns of $\widehat{M}_p^{(k+1)}$ while setting $\widehat{\lambda}^{(k+1)}$ to those norms
12: $k=k+1$
13: **end**
14: $\widehat{\Upsilon}=\bar{\mathcal{Y}}-\langle\bar{\mathcal{X}}|\widehat{\mathcal{B}}_{CP}\rangle$

OP format. For an OP-formatted $\mathcal{B}$, i.e.,

$$\mathcal{B}_{OP}=\circ[\![M_1,\ldots,M_p]\!], \tag{34}$$

we use Theorem 2.1(b) to express (15) as

$$\mathcal{Y}_i=[\![\mathcal{X}_i;M_1,\ldots,M_p]\!]+\mathcal{E}_i. \tag{35}$$

We estimate the parameters in (35) by applying the kth mode matricization for each $k=1,\ldots,p$ on both sides as $\mathcal{Y}_{i(k)}=M_kG_{ik}^{OP}+E_i$, where $G_{ik}^{OP}=\mathcal{X}_{i(k)}(\bigotimes_{j=p,j\neq k}^{1}M_j')$ and $E_i\overset{iid}{\sim}\mathcal{N}_{[m_k,m_{-k}]'}(0,\sigma^2\Sigma_k,\Sigma_{-k})$. Given the similarities between this formulation and (29), the MLEs of the factor matrices are as in (33) but with G_{ik}^{OP} instead of G_{ik}^{CP}. The optimization procedure is similar to Algorithm 2, with the difference again that $M_1,\ldots,M_{p-1}$ are normalized to have unit Frobenius norm. We conclude by noting that (35) is the multilinear tensor regression setup of [18], and for $p=2$ is the matrix-variate regression framework of [45] and [46]. So the OP format frames existing methodology within the ToTR framework.

TR format. Let $\mathcal{B}$ have TR format (11), i.e.,

$$\mathcal{B}_{TR}=\mathrm{tr}(\mathcal{L}_1\times^1\mathcal{L}_2\times^1\ldots\times^1\mathcal{L}_l\times^1\mathcal{M}_1\times^1\mathcal{M}_2\times^1\ldots\times^1\mathcal{M}_p), \tag{36}$$

of TR rank $(s_1,\ldots,s_l,g_1,\ldots,g_p)$, where $\mathcal{L}_j$ and $\mathcal{M}_k$ are third order tensor of sizes $(s_{j-1}\times h_j\times s_j)$ and $(g_{k-1}\times m_k\times g_k)$ respectively, for all $j=1,\ldots,l$ and $k=1,\ldots,p$, and where $g_0=s_l$ and $s_0=g_p$. The TR format reduces the number of unconstrained parameters in $\mathcal{B}$ from $\prod_{i=1}^{l}h_i\prod_{i=1}^{p}m_i$ to $\sum_{j=1}^{l}s_{j-1}h_js_j+\sum_{k=1}^{p}g_{k-1}m_kg_k$. To estimate parameters, we apply the kth mode matricization for $k=1,\ldots,p$ on both sides of (13), yielding

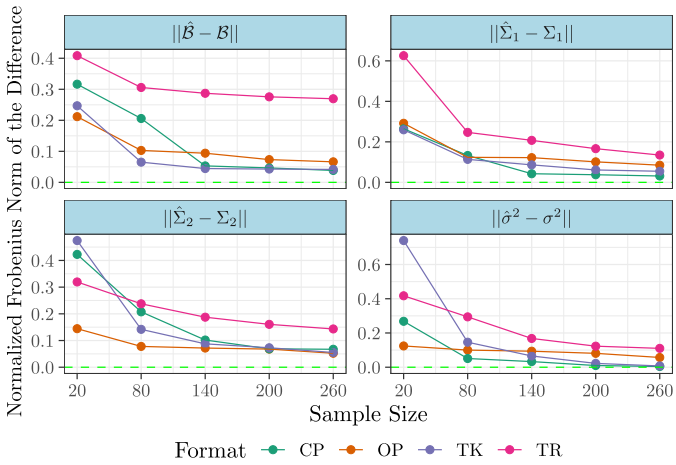


Fig. 6. Performance of the four models presented in Section 3.1, with each model corresponding to a different format on $\mathcal{B}$. Each plot corresponds to the normalized Frobenius norm of the difference between the estimated and true population parameters. We observe that in all cases, an increase in sample size corresponds to more accurate estimates.

$$\mathcal{Y}_{i(k)}=\mathcal{M}_{k(2)}G_{ik}^{TR}+E_i,\quad E_i\overset{iid}{\sim}\mathcal{N}_{[m_k,m_{-k}]'}(0,\sigma^2\Sigma_k,\Sigma_{-k}), \tag{37}$$

and the vectorization for $k=1,\ldots,l$, which gives us

$$\mathrm{vec}(\mathcal{Y}_i)=H_{ik}^{TR}\mathrm{vec}(\mathcal{L}_k)+\boldsymbol{e}_i,\quad \boldsymbol{e}_i\overset{iid}{\sim}\mathcal{N}_m(0,\sigma^2\Sigma), \tag{38}$$

where G_{ik}^{TR} and H_{ik}^{TR} are matrices as defined in Section S1.5.2, available online. Fig. 5 represents tensor-variate versions of H_{ik}^{TR} and G_{ik}^{TR} for when $p=l=3$. Because (29) and (32) are similar to (37) and (38), our ML estimators mirror the CP format case but by replacing $(H_{ik}^{CP},G_{ik}^{CP})$ with $(H_{ik}^{TR},G_{ik}^{TR})$. In this case, estimating $(M_k,\mathrm{vec}(L_k))$ corresponds to estimating $(\mathcal{M}_{k(2)},\mathrm{vec}(\mathcal{L}_k))$. The optimization procedure is similar to Algorithm 2, with the difference in this case being that each factor tensor, other than $\mathcal{M}_p$, is scaled to have unit Frobenius norm. We end here by noting that the special TR case of TT format has been used for ToTR in [22], with $\Sigma=I$, and hence $\Sigma_k=I$ for all k.

Concluding Remarks

Fig. 5 summarizes our estimation methods for $\mathcal{B}$ of different formats. We see that in many cases, an algorithmic block can be made to correspond to a linear model by appropriate choice of $\mathcal{G}_{ik}$ or $\mathcal{H}_{ik}$. Then, fitting a tensor-response linear model involves sequentially fitting smaller-dimensional linear models (one for each tensor factor) until convergence. This intuition behind the estimation of $\mathcal{B}$ is not restricted to the TK, CP, OP and TR formats, but can also help guide estimation algorithms for other formats such as the hierarchical Tucker and the tensor tree formats [47], [48].

While the OP format has the advantage of parsimony, it does not allow for the level of recovery to be adjusted through a tensor rank (as will be illustrated in Section 3 and Fig. 6). The CP format is a more attractive alternative, since it is the natural generalization of the low-rank matrix format (a tensor with CP rank k is the sum of k tensors with CP rank 1). However, the CP format can be too restrictive in some scenarios, such as when the tensor modes have very different sizes. In such cases, the Tucker format provides a more appealing alternative. Moreover, a Tucker-formatted tensor

has the interpretation of being a core tensor that is stretched on each mode by a tall matrix. However, its disadvantage is that the core tensor has the same number of modes as the original tensor, making it impractical in very high-order tensor scenarios. In such situations, the TR format is preferred because each additional tensor-mode requires only one additional tensor factor.

2.3.3 Estimation of σ^2

In all cases, the estimation of Σ_k involves finding the sum of squared errors along the k-mode S_k, as in Equations (26) and (30). Given the estimated $\widehat{\Sigma}_k$ and S_k, the estimate of σ^2 is very cheap and given as

$$\widehat{\sigma}^2 = \frac{1}{nm}\mathrm{tr}(\widehat{\Sigma}_k^{-1} S_k). \tag{39}$$

Thus $\widehat{\sigma}^2$ is obtained alternatingly within each iteration. Moreover, the log-likelihood function evaluated at the current estimated values is greatly simplified based on $\widehat{\sigma}^2$ as

$$\ell = -\frac{nm}{2}\left[1 + \log(2\pi\widehat{\sigma}^2) + \sum_{k=1}^{p} \log|\widehat{\Sigma}_k|/m_k\right].$$

2.3.4 Initialization and Convergence

For local optimality, we need the two conditions that the log-likelihood ℓ is jointly continuous and that the set $\{\boldsymbol{\theta} : \ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}^{(0)})\}$, for a set of initial values $\boldsymbol{\theta}^{(0)}$, is compact [42]. These conditions are satisfied because of the TVN distributional assumption on our errors, as long as the initial values satisfy the constraints on the parameters. We initialized the tensor factor entries in $\mathcal{B}$ with draws from the $\mathcal{U}(0,1)$ distribution. With the TK format, M_k has the constraint $M_k'\Sigma_k^{-1}M_k = I_{d_k}$ for $k = 1,\ldots,p$, so we used $\Sigma_k^{\frac{1}{2}}U$ as its initializer, with U having the left singular vectors of a random matrix of the same order as M_k. We also suggest using identity matrices to initialize $\Sigma_1,\ldots,\Sigma_p$ and σ^2 can be initialized with 1. Our algorithms are declared to converge when we have negligible changes in the loglikelihood, as simplified in Section 2.3.3. A different criteria is the difference in norm $||\mathcal{B}|| + ||\sigma^2\Sigma||$, where $||\sigma^2\Sigma|| = \sigma\prod_{k=1}^{p}||\Sigma_k||$ and $||\mathcal{B}||$ simplifies as per format:

- for $\mathcal{B}_{TK}$, with A^q as the Q matrix from the LQ decomposition of A [33], we have $||\mathcal{B}_{TK}|| = |||[\![\mathcal{V}; L_1^q,\ldots,L_l^q, M_1^q,\ldots,M_p^q]\!]|||$.
- for $\mathcal{B}_{CP}$, $||\mathcal{B}_{CP}||^2 = \sum_{k,l=1}^{R}\{diag(\boldsymbol{\lambda})[*_{i=1}^{l}(L_i'L_i)] * [*_{i=1}^{p}(M_i'M_i)]\}(k,l)$, where "$*$" is the Hadamard, or entry-wise product [33].
- for $\mathcal{B}_{OP}$, $||\mathcal{B}_{OP}|| = \prod_{i=1}^{p}||M_i||$.

2.4 Properties of Our Estimators

2.4.1 Computational Complexity

We derive the computational complexity of our estimation algorithms. Recall that in all cases the response $\mathcal{Y}_i \in \mathbb{R}^{\times_{k=1}^{p} m_k}$ and the covariate $\mathcal{X}_i \in \mathbb{R}^{\times_{k=1}^{l} h_k}$ where l and p are considered fixed. WLOG, we assume that $m_1 = \max\{m_1,\ldots,m_p\}$ and $h_1 = \max\{h_1,\ldots,h_l\}$.

Theorem 2.3. *The computational complexity of our ToTR algorithms when $\mathcal{B}$ has*

1) *the TK format of Section 2.3.2.1, with $d = \prod_{q=1}^{p} d_q$, $d_{-1} = d/d_1$, $c = \prod_{q=1}^{l} c_q$, $d_1 = \max\{d_1,\ldots,d_p\}$ and $c_1 = \max\{c_1,\ldots,c_l\}$ and implemented in Algorithm 1 is $\mathcal{O}(nhc_1 + n^2c + n^2m_1d_{-1} + nm_1^2d_{-1}) + \mathcal{O}(nmd_1) + \mathcal{O}(ncdh_1 + nc_1^2h_1^2d + h_1^3c_1^3) + \mathcal{O}(m_1^3 + nmm_1)$.*
2) *the CP format of Section 2.3.2.2 and implemented in Algorithm 2 is $\mathcal{O}(nh_1^2r^2 + nrm + rm_1^2 + m_1^3 + h_1^3r^3 + nrh + m_1r^2 + nmm_1)$.*
3) *the TR format, as described in Section 2.3.2.4 and with $g_0g_1 = \max\{g_{k-1}g_k : k = 1,2,\ldots p\}$, $s_0s_1 = \max\{s_{k-1}s_k : k = 1,2,\ldots l\}$, $g = \max\{g_0,g_1\}$, $g_0g_1 \leq m_1$ is $\mathcal{O}(mg_1g_0g_p + hnss_l + mh_1^2s_0^2s_1^2 + h_1^3s_0^3s_1^3) + \mathcal{O}(hs_1s_0s_l + m_1^3 + g_0^3g_1^3 + nmm_1)$.*

Proof. See Sections S1.7.1 - S1.7.4 for the proofs, available online. □

In all cases we have the term $\mathcal{O}(nmm_1)$, which is the complexity of obtaining the sum of square errors S_k of Equation (26) across the largest tensor-response mode, and it is necessary for obtaining the scale matrices $\Sigma_1,\ldots,\Sigma_p$. In many cases this term will dominate the computational complexity. However, $O(nmm_1)$ is considerably smaller than $O(nm^2)$, which would be the case where our complexity increases quadratically with the dimensionality of the tensor response. We also note that in all cases, the cubic terms are WRT the tensor ranks, which can be considered negligible because such ranks are often chosen to be small, in the spirit of scientific parsimony. Finally, for the TK format, some of the factors can assumed to be identity matrices, allowing us to further reduce the complexity. (See Section S1.7.3, available online, for the computational complexity of the OP format under specific conditions.)

2.4.2 Asymptotic Sampling Distributions

We now derive the asymptotic distributions of our model-estimated parameters, specifically, the linear component and the covariance component (Section S1.13), available online.

We first explore the limiting distribution of the estimated linear components $\mathrm{vec}(\widehat{\mathcal{B}})$, which in all cases is multivariate normal with mean $\mathrm{vec}(\mathcal{B})$ satisfying the same low-rank format of $\mathrm{vec}(\widehat{\mathcal{B}})$. For the Tucker format, we first show that the vectorized core tensor $\mathrm{vec}(\widehat{\mathcal{V}})$ follows a non-singular multivariate normal distribution, and therefore by Slutsky's theorem $\mathrm{vec}(\widehat{\mathcal{B}})$ follows a singular multivariate normal distribution, where the singularity of the covariance matrix constraints the limiting distribution to the original low-rank Tucker format. For the CP, TR and OP cases we first show that the low-rank format factors in $\mathcal{B}$ are jointly normally distributed. Therefore, by the Delta method the estimated tensors $\widehat{\mathcal{B}}$ in vectorized form are asymptotic normally distributed. In this case the resulting multivariate normal distribution is also singular, but these are only approximations to the CP, TR or OP formats and not constraints on the limiting distributions like it is the case for the Tucker format.

For the remainder of this paper, we define $h \doteq \prod_{i=1}^{l} h_i$, $M \doteq \bigotimes_{i=p}^{1} M_i$ and $L \doteq \bigotimes_{i=l}^{1} L_i$. We first assume that (15) holds without an intercept.

Theorem 2.4. *Let (15) hold with $\mathcal{B} \equiv \mathcal{B}_{TK}$ of Tucker format as in (18) and let $X = [vec(\mathcal{X}_1) \ldots vec(\mathcal{X}_n)]$ and $\widehat{\mathcal{B}}_{TK} = [\![\widehat{\mathcal{V}}; \widehat{L}_1, \widehat{L}_2, \ldots, \widehat{L}_l, \widehat{M}_1, \widehat{M}_2, \ldots, \widehat{M}_p]\!]$. Then as $n \to \infty$*

$$\mathrm{vec}(\widehat{\mathcal{B}}_{TK}) \xrightarrow{d} \mathcal{N}_{mh}\Big(\mathrm{vec}(\mathcal{B}_{TK}), \sigma^2(MM') \otimes (P_L(XX')^{-1}P_L)\Big),$$

where $P_L = \bigotimes_{i=l}^{1} P_i$ and $P_i = L_i(L_i'L_i)^{-1}L_i'$.

Proof. See Section S1.8, available online. □

The limiting distribution in Theorem 2.4 is TVN when XX' has a Kronecker structure: examples include factorial designs and B-splines [49]. Here we present one such case.

Corollary 2.1. *When $\widehat{\mathcal{B}}_{TK}$ is used to estimate a balanced TANOVA with q units for each factor combination, then for $\boldsymbol{s} = [h_1, h_2, \ldots, h_l, m_1, m_2, \ldots, m_p]'$, as $n \to \infty$*

$$\widehat{\mathcal{B}}_{TK} \xrightarrow{d} \mathcal{N}_{\boldsymbol{s}}\Big(\mathcal{B}, \frac{\sigma^2}{q} P_1, P_2, \ldots, P_l, M_1M_1', M_2M_2' \ldots, M_pM_p'\Big).$$

Proof. Here $XX' = qI_h$ and so the variance-covariance matrix in the limiting distribution of Theorem 2.4 is $(\sigma^2/q)(MM') \otimes (\bigotimes_{i=l}^{1} P_i)$, which is Kronecker-separable. The result follows from Definition 2.2. □

The CP format case is similar to Theorem 2.4.

Theorem 2.5. *Consider (15) with $\mathcal{B} \equiv \mathcal{B}_{CP}$ as in (28) and the ML estimator $\widehat{\mathcal{B}}_{CP} = [\![\widehat{L}_1, \widehat{L}_2, \ldots, \widehat{L}_l, \widehat{M}_1, \widehat{M}_2, \ldots, \widehat{M}_p]\!]$. Then*

$$\mathrm{vec}(\widehat{\mathcal{B}}_{CP}) \xrightarrow{d} \mathcal{N}_{mh}\big(\mathrm{vec}(\mathcal{B}_{CP}), J_{CP}R_{CP}(\mathrm{I}_n \otimes \Sigma)R_{CP}'J_{CP}'\big)$$

as $n \to \infty$. Here J_{CP} is a Jacobian matrix and is given along with the block matrix R_{CP} in Section S1.9, available online.

Proof. See Section S1.9, available online. □

The sampling distributions of $\mathcal{B}$ under the OP or TR formats are similar to the CP case, and are in theorem S1.2 of Section S1.10, available online.

Theorem 2.6. *For a model with intercept, as in (13), Theorems 2.4 and 2.5 also hold after centering the covariates.*

Proof. See Section S1.12, available online. □

Section S1.13, available online, also discusses inference on the scale components $\widehat{\Sigma}_1, \widehat{\Sigma}_2, \ldots, \widehat{\Sigma}_p$. Theorem S1.3, available online, establishes the asymptotic independence of the scale and the linear components, we find the Fisher information matrix WRT the scale parameters, and establish its singularity. Our results on the asymptotic distribution of the scale components are not unique to our regression methodology but also generally hold for the TVN distribution.

2.5 Model Selection or Rank Determination

For the CP, TR and TK formats, we determine optimal ranks using the Bayesian information criterion (BIC) [50], [51]. This calculation requires the loglikelihood that we simplified in Section 2.3.3. Section S1.6, available online, provides more details on rank determination (equivalently, model selection) and on the total number of calculations needed to obtain the BIC.

3 Experimental Evaluations

We study estimation performance of the scale parameters and the low-rank linear component $\mathcal{B}$ using simulation experiments on different ToTR models. Section 3.1 assesses the consistency of our estimators, and Section 3.2 evaluates the amounts of recovery that different low-rank formats have of $\mathcal{B}$, and the impact of noise on discrimination in a TANOVA framework.

3.1 A TANOVA(2,2) Model With Low-Rank Formats

We simulated observations from the matrix-on-matrix regression (MoMR) model

$$Y_{ijk} = \langle X_{ij}|\mathcal{B}\rangle + E_{ijk}, \quad E_{ijk} \overset{iid}{\sim} \mathcal{N}_{[6,7]'}(0, \sigma^2\Sigma_1, \Sigma_2), \tag{40}$$

where $i = 1, 2, 3, 4$, $j = 1, 2, 3, 4, 5$ and X_{ij} is a 4×5 matrix with 1 at the (i, j)th position and zeroes everywhere else. We set $\sigma^2 = 1$ and obtained Σ_1 and Σ_2 independently from Wishart distributions, that is, we obtained $\Sigma_1 \sim \mathcal{W}_6(6, I_6)$ and $\Sigma_2 \sim \mathcal{W}_7(7, I_7)$ before scaling each by their (1,1)th element. We obtained realizations from four MoMR models, one each for $\mathcal{B}$ of TK, CP, TR and OP formats, and fit appropriate models to the data using the ML estimation procedures described in Section 2.3. To study consistency properties of our estimators, we used $k = 1, 4, 7, 10$ and 13, meaning that our sample sizes ranged over $n \in \{20, 80, 140, 200, 260\}$. An unstructured $\mathcal{B}$ in this experiment would have $4 \times 5 \times 6 \times 7 = 840$ entries, but $\mathcal{B}$ has only 59 unconstrained parameters with the OP format and 45 unconstrained parameters when it is of CP format of rank 2. This number is only 60 when $\mathcal{B}$ has TK format with rank (2,2,2,2) and 70 when it is of the TR format with rank (2,2,2,2). Thus, our lower-rank simulation framework had at least 91% fewer unconstrained parameters in $\mathcal{B}$. We simulated data from (40) using $\mathcal{B}$, σ^2, Σ_1 and Σ_2 and estimated the parameters for each replication. Fig. 6 displays the Frobenius norm of the difference between the true and estimated parameters, and shows that as sample size increases, $(\widehat{\mathcal{B}}, \widehat{\sigma}^2, \widehat{\Sigma}_1, \widehat{\Sigma}_2)$ approach the true parameters $(\mathcal{B}, \sigma^2, \Sigma_1, \Sigma_2)$, demonstrating consistency of the estimators.

3.2 Evaluating Recovery and Discrimination

We simulated 600 observations from the MoMR model

$$Y_{ijk} = \langle X_{ij}|\mathcal{B}\rangle + E_{ijk}, \quad E_{ijk} \overset{iid}{\sim} \mathcal{N}_{[87,106]'}(0, \sigma^2\Sigma_1, \Sigma_2), \tag{41}$$

where $i = 1, 2, 3, 4$, $j = 1, 2, 3$, X_{ij} is a 4×3 matrix with 1 at the (i, j)th position and zero elsewhere, and $\Upsilon_{ij} = \langle X_{ij}|\mathcal{B}\rangle$ corresponds to the pixel-wise logit transformation of the jth additive color (Red, Green, Blue) of the ith Andean camelid (Guanaco, Llama, Vicuña, Alpaca) images of Fig. 7a. We set Σ_1 and Σ_2 to be AR(1) correlation matrices with coefficients 0.1 and -0.1 respectively, and $\sigma^2 = 1$. Sans constaints, we have $3 \times 4 \times 87 \times 106 = 110664$ parameters in $\mathcal{B}$.

We fit (41) separately for TR-, TK-, CP- and OP-formatted $\mathcal{B}$, with ranks set to have similar number of unconstrained parameters in $\mathcal{B}$. The TR format with rank (3,3,5,3) had 2958 such parameters, the TK format with rank (4,4,9,9) had 3061 parameters, the CP format with rank 15 had 3001 parameters in $\mathcal{B}$, while the relatively inflexible OP format had 666

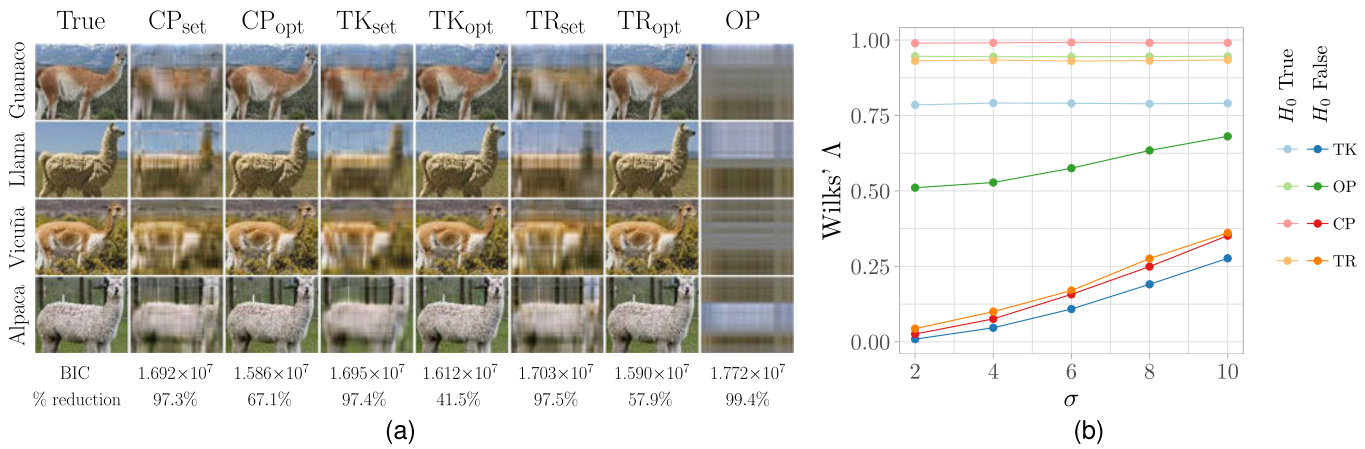


Fig. 7. (a) Results of fitting seven TANOVA(2,2) models on data simulated from (41). One model was fit assuming the OP format, and two models were fit for each of the TK, CP and TR formats: one with set ranks and another one with optimal ranks, as chosen by BIC. The two factors are the type of andean camelid (Guanaco, Llama, Vicuña, Alpaca) and the type of additive color RGB (red, green, blue). The Guanaco, Llama and Alpaca images are from Wikipedia Commons and the Vicuña image is from Encyclopædia Britannica. It is evident that while one cannot adjust the OP rank, increased rank for the TR, TK and CP formats result in more image restoration. (b) Monte-Carlo 95th quantiles of the Wilks' Λ statistics that test the set of hypothesis in equation (42), for the OP, CP, TR and TK formats, five values of scalar variance σ^2 in the x-axis, and for both true and false null hypotheses. Large variabilities lead to larger test statistics when H_0 is false, leading to weaker evidence against the null hypothesis. In all cases, CP, TR and TK formatting leads to more significant statistics when compared to the OP, even when the ranks are not optimal.

parameters. In all cases, the dimension of $\mathcal{B}$ was reduced by over 97%. The estimated tensor $\mathcal{B}$ in each case corresponds to the estimated color images of the four Andean camelids. Fig. 7a shows varying success of these four formats in recovering the underlying camelid image (true $\mathcal{B}$). The OP-estimated images are the least-resolved, with the reduced number of parameters for $\mathcal{B}$ inadequate for recovery. But the other formats can adjust for the quantum of reduction in parameters through their ranks. We illustrate this aspect by fitting (41) with $\mathcal{B}$ having the TK, CP and TR formats with optimal rank chosen by BIC, following Section 2.5. Fig. 7a shows very good recovery of $\mathcal{B}$ by these BIC-optimized $\widehat{\mathcal{B}}$s, with unappreciable visual differences in all cases. In contrast, the model fit with unstructured $\mathcal{B}$ and diagonal Σ, has a BIC of 1.64×10^7, while fitting a model with a similar $\mathcal{B}$ but Kronecker-separable Σ has a BIC of 1.63×10^7. The CP, TK and TR formats therefore outperform these two alternatives when the ranks are tuned.

The TANOVA(2,2) formulation of (41) enables us to test

$$\begin{aligned} &\mathrm{H}_0 : \mathcal{P}_1 = \mathcal{P}_2 = \mathcal{P}_3 = \mathcal{P}_4 \qquad \text{vs.} \\ &\mathrm{H}_\mathrm{a} : \mathcal{P}_i \neq \mathcal{P}_{i^*}, \text{ for some } i \neq i^* \in \{1,2,3,4\} \end{aligned} \tag{42}$$

where $\mathcal{P}_i$ is a third-order tensor of size $3 \times 87 \times 106$ that contains the RGB slices of the ith Andean camelid image. The usual Wilks' Λ statistic [52] is $\Lambda = |\widehat{\Sigma}_R|/|\widehat{\Sigma}_T|$, where $\widehat{\Sigma}_R$ is the sample covariance matrix of the residuals and $\widehat{\Sigma}_T$ is the sample covariance matrix of the simpler model's residuals, which finds a common mean across all camelids. (Section S2, available online, details the calculation of $\widehat{\Sigma}_R$ and $\widehat{\Sigma}_T$.) We illustrate the role of σ^2 and the low-rank OP, TR, CP or TK formats in distinguishing the four camelids, as measured by the Wilks' Λ test statistic, in Fig. 7b, for $\sigma = 2, 4, 6, 8, 10$. The value of Λ increases with σ^2, meaning that larger variances decrease the power of our test. Further, the CP, TR and TK formats yield lower-valued (more significant) test statistics than OP. This finding illustrates the limits of the less-flexible OP format relative to the others in recovering $\mathcal{B}$. Nevertheless, OP joins the other three formats in consistency of estimation and discrimination, as illustrated by Wilks' Λ.

4 Real Data Applications

Having evaluated performance of our reduced-rank ToTR methodology, we apply it to the datasets of Section 1.1.

4.1 A TANOVA(1,5) Model for Cerebral Activity

Section 1.1.1 laid out a TANOVA model involving 30 fMRI volumes of voxel-wise changes in activation from a baseline, each volume corresponding to one of ten words connoting death, positive or negative affects, for each of 17 subjects. For the jth subject we have a fifth-order tensor $\mathcal{Y}_j$ of order $3 \times 10 \times 43 \times 56 \times 20$, where the first two modes correspond to the three kinds of word stimulus and the individual words, and the other modes correspond to the dimensions of the image volume. The jth subject has status given by $\boldsymbol{x}_j$ that is a 2D unit vector with 1 at position i that is 1 for attempter or 2 for ideator. We model these responses and covariates as

$$\mathcal{Y}_j = \langle \boldsymbol{x}_j | \mathcal{B} \rangle + \mathcal{E}_j, \quad \mathcal{E}_j \overset{iid}{\sim} \mathcal{N}_{\boldsymbol{m}_1}(0, \sigma^2\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4, \Sigma_5),$$

where $\boldsymbol{m}_1 = [3, 10, 43, 56, 20]'$ and $j = 1, \ldots, 17$. We let $\mathcal{B}$ have the TK format $[\![\mathcal{V}; L_1, M_1, M_2, M_3, M_4, M_5]\!]$ with rank (2,3,6,15,20,7) chosen by BIC from 256 candidate ranks, and where $M_k' \Sigma_k^{-1} M_k = I_{d_k}, k = 1, 2, 3, 4, 5$. The 77578 parameters to be estimated in our $\mathcal{B}$ represent an over 97.3% reduction over that of the unconstrained $\mathcal{B}$ of size $2 \times 3 \times 10 \times 43 \times 56 \times 20$, or 2889600 parameters. (Our use of a TK format exploits its nicer distributional properties for easier inference, and therefore we only use this format here.) We set Σ_1 (specifying relationships between word types) to be unconstrained, Σ_2 (covariances between same kinds of words) to have an equicorrelation structure and $\Sigma_3, \Sigma_4, \Sigma_5$ with AR(1) correlations to capture spatial context in the image volume. Fitting the model with unstructured $\mathcal{B}$ and diagonal Σ yielded a BIC of 210 million, while the fitted model with a similar $\mathcal{B}$ but Kronecker-separable Σ reported a BIC of 164 million. In

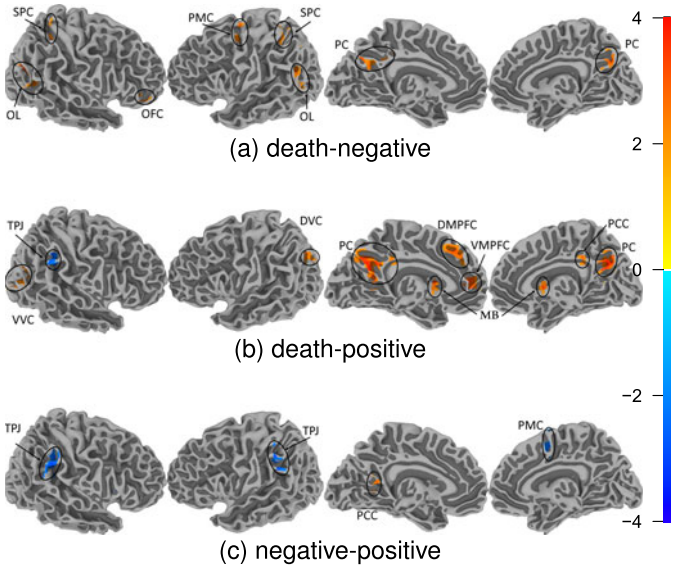


Fig. 8. The test statistic $\widehat{\mathcal{Z}}_*$ of the interaction between subject's attempter/ideator status and (a) death-negative, (b) death-positive and (c) negative-positive words at voxels identified significant by cluster thresholding at the 5% level. These voxels are in the precuneus (PC), temporal-parietal junction (TPJ), orbital frontal cortex (OFC), premotor cortex (PMC), superior parietal cortex (SPC), ventral visual cortex (VVC), dorsal visual cortex (DVC), dorsal medial frontal cortex (DMPFC), ventral medial prefrontal cortex (VMPFC), mamillary bodies (MB), posterior cingulate cortex (PCC) and occipital lobe (OL).

contrast, our TK model with Kronecker-separable covariance outperformed these two alternatives with a BIC of 127 million.

Our primary interest here is to find regions of significant interaction between word type and subject suicide attempter/ideator status to determine markers for suicide risk assessment and intervention. The interaction estimate can be expressed as $\widehat{\mathcal{B}}_* = \widehat{\mathcal{B}} \times_1 \boldsymbol{c}_1' \times_2 C_2 \times_3 \boldsymbol{c}_3'$, where $\boldsymbol{c}_1$ is a contrast vector that finds differences between suicide attempter/ideation status, C_2 is a contrast matrix for differences across word type and $\boldsymbol{c}_3$ is a contrast vector that averages the ten words of each type. These contrast matrices and vectors are given in (S30), available online. From Theorem 2.4,

$$\widehat{\mathcal{B}}_* \xrightarrow{d} \mathcal{N}_{\boldsymbol{m}_2}(\mathcal{B}_*, \tau^2 C_2 M_1 M_1' C_2', M_3 M_3', M_4 M_4', M_5 M_5'),$$

where $\boldsymbol{m}_2 = [3, 20, 43, 56]'$, $\mathcal{B}_* = \mathcal{B} \times_1 \boldsymbol{c}_1' \times_2 C_2 \times_3 \boldsymbol{c}_3'$ and τ^2 is as in Section S3.1, available online. Using the asymptotic distribution of $\widehat{\mathcal{B}}_*$, we marginally standardize it to obtain $\widehat{\mathcal{Z}}_*$ as shown in (S33), available online, which also follows the TVN distribution but with correlation matrices as scale parameters. In Section S3.1, available online, we detail its derivation, interpretation and asymptotic distribution. For the ith level interaction, consider the set of hypotheses at the (k, l, m)th voxel

$$H_o : \mathcal{B}_*(i, k, l, m) = 0 \quad \text{vs} \quad H_a : \mathcal{B}_*(i, k, l, m) \neq 0.$$

Under the null hypothesis of no ith interaction effect at the (k, l, m)th voxel, the marginal distribution of $\widehat{\mathcal{Z}}_*(i, k, l, m)$ is asymptotically $N(0, 1)$. Fig. 8 displays 3D maps of the brain with significant values of $\widehat{\mathcal{Z}}_*$ overlaid for each of the three pairs of interactions. Significant voxels were decided using cluster thresholding [53] ($\alpha = 0.05$), with clusters of at least 12 contiguous (under a second-order neighborhood specification) voxels, with this minimum cluster size determined by the Analysis for Neuroimaging (AFNI) software [54], [55]. There are many methods [56], [57], [58], [59], [60], [61], [62] for significance detection in fMRI studies but we use cluster thresholding here as an illustration and also because it is the most popular method. We now briefly discuss the results.

Fig. 8a identifies significant interactions between death- and negative-connoting words on the one hand and suicide attempters vis-a-vis ideators on the other. All significant interactions are positive and dominated by the precuneus and the orbital frontal cortex. The precuneus is associated with depression and rumination [63], [64], [65], while the orbital frontal cortex is associated with the influence that emotions and feelings have on decision-making [66], as well as with suicide attempters' reactions to external stimuli [67]. Both regions are also associated with the Default Mode Network (DMN) that plays a role in representing emotions [68]. These results indicate more differential rumination and emotions (between attempters and ideators) caused by death-related words, as compared to negative-connoting words. These findings are reinforced by the significance detected in the occipital lobe, the premotor cortex (PMC) and the superior parietal cortical regions that are related to working memory and depression [69], [70], [71]. Fig. 8b displays significant interactions between the positive and death-related words and suicide attempters and ideators. The precuneus is more pronounced here relative to Fig. 8a, indicating that death-related words are more salient than words that have negative and positive connotations among attempters vis-a-vis ideators. This observation is reinforced with the detected significance in the dorsal and ventral visual medial prefrontal cortex, the mammillary bodies, and the posterior cingulate cortex (PCC) that are all involved in processing emotional information [72], [73]. The PCC is also involved in memory, emotion, and decision-making [74], [75] and is connected to the temporal-parietal junction [76] which is involved with emotions and perception [77], [78]. High $\widehat{\mathcal{Z}}_*$ values in the ventral and dorsal visual cortices are commensurate with their association with working memory tasks [79]. Also, the low values of $\widehat{\mathcal{Z}}_*$ in the temporal parietal junction point to needed additional processing of death-related versus positive-connoting words among attempters relative to ideators. Fig. 8c shows significant interactions between negative and positive-emoting words and suicide attempters and ideators. The low values in the left and right temporal-parietal junctions and the PMC indicate that words conveying negative thoughts don't need as much processing as do positive-connoting words among attempters relative to ideators. The significant association of the PCC in both Figs. 8b and 8c supports our hypothesis that death-related words are more salient than negative or positive words in differentiating attempters from ideators. In summary, the two groups of subjects have positive- and negative-connoting words result in neurally similar significant brain regions when compared to death-related words, which show further significance in areas associated with the processing of emotional feelings and planning. Our conclusions here are on an experiment with only 9 attempters and 8 ideators and so are preliminary, but are intrepretable, providing some confidence in the

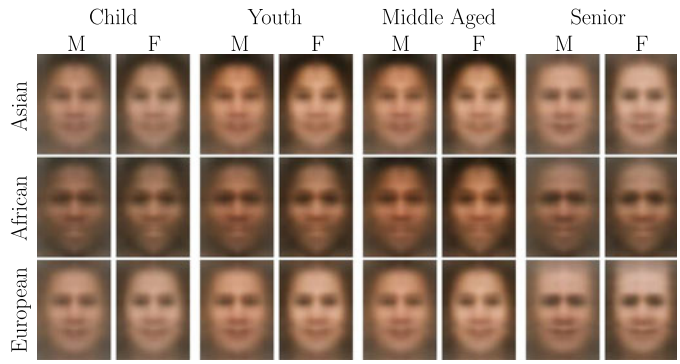


Fig. 9. Different slices of the resulting factorized tensor $\mathcal{B}$ that results from fitting a TANOVA(3,3) model on the LFW dataset using the TT format. The results are compressed mean images across genders (male, female), ethnic origin (Asian, African, European) and age groups (child, youth, middle aged and senior) from 605 central LFW images. We can observe that the TT format preserved vital information regarding the factor-combination of age group, gender and ethnic-origin.

practical reductions afforded by TANOVA when coupled with the use of the Tucker-formatted $\mathcal{B}$ for this application.

4.2 A TANOVA(3,3) Model for the LFW Face Database

We return to the LFW database of Section 1.1.2 that is a compendium of over 13,000 face images. Using the steps detailed in Section S3.2, available online, we selected 605 images with unambiguous genders, age group and ethnic origin, and such that there are at most 33 images for each factor combination. This dataset was also used by [19] with the goal of classification, leading to a vector-variate response of attributes and tensor-valued covariates of color images, for which a CP format was assumed. In contrast, our objective is to distinguish the characteristics of different attributes, leading to a TANOVA(3,3) model with color images as the response and gender, ethnic origin and gender as covariates. Our model is as per (15) and specifically

$$\mathcal{Y}_{ijkl} = \langle \mathcal{X}_{ijk} | \mathcal{B} \rangle + \mathcal{E}_{ijkl}, \quad \mathcal{E}_{ijkl} \sim \mathcal{N}_{\boldsymbol{m}_3}(0, \sigma^2 \Sigma_1, \Sigma_2, \Sigma_3),$$

where $i = 1, 2, \quad j = 1, 2, 3, k = 1, 2, 3, 4, l = 1, \ldots, n_{ijk}$, $\boldsymbol{m}_3 = [151, 111, 3]'$, and the (i, j, k, l)th response $\mathcal{Y}_{ijkl}$ is the color image of size $151 \times 111 \times 3$ for the lth person of the ith gender, jth ethnic origin and kth age group. Here $\mathcal{X}_{ijk} = \boldsymbol{e}_i^2 \circ \boldsymbol{e}_j^3 \circ \boldsymbol{e}_k^4$ is the tensor-valued covariate for a TANOVA(3,3) model with $(h_1, h_2, h_3) = (2, 3, 4)$, as described in Section 2.2.2, encoding the genders $\times$ ethnic-origin $\times$ age-group attributes of $\mathcal{Y}_{ijkl}$. The corresponding TANOVA parameter $\mathcal{B}$ is of size $2 \times 3 \times 4 \times 151 \times 111 \times 3$ and contains all the group means. We constrained $\mathcal{B}$ to have a tensor train (TT) format of TR rank (1,3,3,4,10,3), chosen using BIC out of a total of 64 candidate ranks. (In terms of the BIC, the TT format also bested the TK, CP and OP formats.) The number of parameters involved in $\mathcal{B}$ is 6393 due to the TT restriction, which is a reduction in the number of unconstrained parameters of around 99% from the unconstrained $\mathcal{B}$ that has more than 1.2 million parameters. Fig. 9 displays the estimated $\widehat{\mathcal{B}}$, from where we observe that the TT format preserved visual information regarding ethnic origin, gender, and age-group. Fitting the model with unstructured $\mathcal{B}$ and diagonal Σ resulted in a BIC of 1.02×10^9, while the fitted model with a similar $\mathcal{B}$ but Kronecker-separable Σ reported a BIC of -1.22×10^7. In contrast, our TT model with Kronecker-separable covariance outperformed these alternatives with a BIC of -1.87×10^7.

5 Discussion

We have provided a multivariate regression and ANOVA framework that exploits the tensor-valued structure of the explanatory and response variables using four different low-rank formats on the regression coefficient and a Kronecker-separable structure on the covariance matrix. These structures are imposed for context but more so for practical reasons, as the number of parameters involved in the classical MVMLR model grows exponentially with the tensor dimensions. Different stuctures can be compared between each other using criteria such as BIC. We provided algorithms for ML estimation, derived their computational complexity, implemented them in an R package (`totr`), and evaluated them via simulation experiments. We also studied the asymptotic properties of our estimators and applied our methodology to identify brain regions associated with suicide attempt or ideation status and death- negative- or positive-connoting words. Finally, we also used our methods to distinguish facial characteristics in the LFW dataset.

A reviewer has asked whether our ToTR methodology, that is based on linear modeling, has advantages over nonlinear deep learning methods in the context of the applications of Section 4. We contend that there are several aspects that make ToTR more suitable than deep learning in both cases. For one, the linearity of the model is dictated by the experimental setup in both cases. Second, deep learning generally requires substantial amounts of training data. For the brain imaging application of Section 4.1, we only have data on 17 subjects, while for the LFW dataset of Section 4.2, we have factor combinations that are also severely imbalanced in sample size, with one factor combination having around 2,000 replications, and some others having only a handful of images. Further, we have developed inference tools for our ToTR methodology which are needed to assess cerebral activity in Section 4.1. That application also shows our ToTR and TANOVA results as easily interpreted. Such benefits are not present with deep learning methodology. Finally, selecting the right deep learning model requires knowledge of the loss function and training method, and calls for considerable skill and expertise. In contrast, our setup allows for a simple objective tool (BIC) for deciding on the optimal model.

There are several other avenues for further investigation. For instance, we can perform additional dimension reduction by adding an L_1 penalty on the likelihood optimization. Also, the number of parameters in the intercept can potentially grow when the tensor response is high-dimensional, which motivates specifying a low-rank structure on the intercept term. Similarly, the independent identically distribution assumption on the errors is not feasible when external factors group data-points into units that are similar to one another. For these cases, a mixed-effects model is more appropriate. Further, it would be worth investigating generalization of the Kronecker separable structure of the dispersion matrix or the normality assumption to incorporate more general distributional forms. Finally, it would be interesting to study the

exact distribution of Wilks' Λ statistic or other statistic that can be used for testing hypothesis in our TANOVA framework without the need to do simulation. These are some issues that may benefit from further attention and that we leave for future work.

ACKNOWLEDGMENTS

The authors are very grateful to an associate editor and four reviewers, all anonymous, whose detailed comments on earlier versions of this manuscript greatly improved its content. We also thank B. Klinedinst and A. Willette of the Program of Neuroscience and the Department of Food Sciences and Human Nutrition at Iowa State University for help with the interpretation of Fig. 8. The content of this paper is however solely the responsibility of the authors and does not represent the official views of the NIJ, the NIBIB, the NIH, the NIFA or the USDA. A version of this article won the first author first place in the 2022 Student Paper Competition of the American Statistical Association (ASA) Section on Statistics in Imaging.

REFERENCES

[1] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. London, U.K.: Pearson, 2008.
[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodol.)*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[3] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
[4] R. D. Cook, B. Li, and F. Chiaromonte, "Envelope models for parsimonious and efficient multivariate liner regression," *Statistica Sinica*, vol. 20, no. 3, pp. 927–960, 2010.
[5] A. Mukherjee and J. Zhu, "Reduced rank ridge regression and its kernel extensions," *Statist. Anal. Data Mining*, vol. 4, no. 6, pp. 612–622, Dec. 2011.
[6] K. A. Busch, J. Fawcett, and D. G. Jacobs, "Clinical correlates of inpatient suicide," *J. Clin. Psychiatry*, vol. 64, pp. 14–19, 2003.
[7] M. A. Just *et al.*, "Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth," *Nature Hum. Behav.*, vol. 1, pp. 911–919, 2017.
[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[9] M. Afifi and A. Abdelhamed, "AFIF4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces," *J. Vis. Commun. Image Representation*, vol. 62, pp. 77–86, 2019.
[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 365–372.
[11] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. Amer. Statist. Assoc.*, vol. 108, no. 502, pp. 540–552, Dec. 2013.
[12] X. Li, H. Zhou, and L. Li, "Tucker tensor regression and neuroimaging analysis," *Statist. Biosci.*, vol. 10, pp. 520–545, Mar. 2018.
[13] Y. Zhou, R. K. W. Wong, and K. He, "Tensor linear regression: Degeneracy and solution," *IEEE Access*, vol. 9, pp. 7775–7788, 2021.
[14] W. W. Sun and L. Li, "STORE: Sparse tensor response regression and neuroimaging analysis," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4908–4944, Jan. 2017.
[15] L. Li and X. Zhang, "Parsimonious tensor response regression," *J. Amer. Statist. Assoc.*, vol. 112, no. 519, pp. 1131–1146, 2017.
[16] G. Rabusseau and H. Kadri, "Low-rank regression with tensor responses," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1875–1883.
[17] R. Guhaniyogi, S. Qamar, and D. B. Dunson, "Bayesian tensor regression," *J. Mach. Learn. Res.*, vol. 18, no. 79, pp. 1–31, 2017.
[18] P. Hoff, "Multilinear tensor regression for longitudinal relational data," *Ann. Appl. Statist.*, vol. 9, pp. 1169–1193, Sep. 2015.
[19] E. Lock, "Tensor-on-tensor regression," *J. Comput. Graphical Statist.*, vol. 27, pp. 638–647, Jan. 2017.
[20] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.
[21] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory" multi-modal factor analysis," *UCLA Work. Papers Phonetics*, vol. 16, pp. 1–84, 1970.
[22] Y. Liu, J. Liu, and C. Zhu, "Low-rank tensor train coefficient array estimation for tensor-on-tensor regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5402–5411, Dec. 2020.
[23] C. Llosa, "Tensor on tensor regression with tensor normal errors and tensor network states on the regression parameter," *Iowa State Univ. Creative Components*, vol. 82, Jan. 2018. [Online]. Available: https://lib.dr.iastate.edu/creativecomponents/82
[24] W. Zhao, K. M. Kendrick, F. Chen, H. Li, and T. Feng, "Neural circuitry involved in quitting after repeated failures: Role of the cingulate and temporal parietal junction," *Sci. Rep.*, vol. 6, no. 1, Apr. 2016, Art. no. 24713.
[25] I. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.
[26] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep. 1966.
[27] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Sep. 2009.
[28] M. Bahri, Y. Panagakis, and S. Zafeiriou, "Robust kronecker-decomposable component analysis for low-rank modeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3372–3381.
[29] P. Hoff, "Separable covariance arrays via the tucker product, with applications to multivariate relational data," *Bayesian Anal.*, vol. 6, no. 2, pp. 179–196, Jun. 2011.
[30] D. Akdemir and A. Gupta, "Array variate random variables with multiway kronecker delta covariance matrix structure," *J. Algebr. Statist.*, vol. 2, pp. 98–113, Jan. 2011.
[31] M. Ohlson, M. R. Ahmad, and D. von Rosen, "The multilinear normal distribution: Introduction and some basic properties," *J. Multivariate Anal.*, vol. 113, pp. 37–47, 2013.
[32] A. M. Manceur and P. Dutilleul, "Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion," *J. Comput. Appl. Math.*, vol. 239, pp. 37–49, 2013.
[33] T. G. Kolda, "Multilinear operators for higher-order decompositions," Sandia Nat. Lab., Albuquerque, NM, Tech. Rep. SAND2006–2081, 923081, Apr. 2006.
[34] A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, and D. P. Mandic, "Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions," *Found. Trends Mach. Learn.*, vol. 9, no. 4/5, pp. 249–429, 2016.
[35] L. De Lathauwer , B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.
[36] D. Gerard and P. Hoff, "A higher-order LQ decomposition for separable covariance models," *Linear Algebra Appl.*, vol. 505, pp. 57–84, 2016.
[37] R. Orús, "A practical introduction to tensor networks: Matrix product states and projected entangled pair states," *Ann. Phys.*, vol. 349, pp. 117–158, 2014.
[38] G. Z. Thompson, R. Maitra, W. Q. Meeker, and A. F. Bastawros, "Classification with the matrix-variate-t distribution," *J. Comput. Graphical Statist.*, vol. 29, no. 3, pp. 668–674, Jul. 2020.
[39] A. Gupta and D. Nagar, *Matrix Variate Distributions*. New York, NY, USA: Taylor & Francis, 1999.
[40] M. M. Wolf, F. Verstraete, M. B. Hastings, and J. I. Cirac, "Area laws in quantum systems: Mutual information and correlations," *Phys. Rev. Lett.*, vol. 100, Mar. 2008, Art. no. 070502.
[41] M. S. Srivastava, T. von Rosen, and D. von Rosen, "Models with a kronecker product covariance structure: Estimation and testing," *Math. Methods Statist.*, vol. 17, no. 4, pp. 357–370, Dec. 2008.
[42] J. de Leeuw, "Block-relaxation algorithms in statistics," in *Information Systems and Data Analysis*, Berlin, Germany: Springer, 1994, pp. 308–324.
[43] H. Abdi, "Singular value decomposition (SVD) and generalized singular value decomposition (GSVD)," *Encyclopedia Meas. Statist.*, vol. 72, no. 2, 2007.
[44] H. Glanz and L. Carvalho, "An expectation maximization algorithm for the matrix normal distribution with an application in remote sensing," *J. Multivariate Anal.*, vol. 167, pp. 31–48, 2018.

[45] C. Viroli, "On matrix-variate regression analysis," *J. Multivariate Anal.*, vol. 111, pp. 296–309, 2012.
[46] S. Ding and R. Cook, "Matrix-variate regressions and envelope models," *J. Roy. Statist. Soc., B (Statist. Methodol,)*, vol. 80, pp. 387–408, May 2016.
[47] W. Hackbusch and S. Kühn, "A new scheme for the tensor representation," *J. Fourier Anal. Appl.*, vol. 15, pp. 706–722, Oct. 2009.
[48] L. Grasedyck, "Hierarchical singular value decomposition of tensors," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 4, pp. 2029–2054, 2010.
[49] I. D. Currie, M. Durban, and P. H. C. Eilers, "Generalized linear array models with applications to multidimensional smoothing," *J. Roy. Statist. Soc. B (Statist. Method.)*, vol. 68, no. 2, pp. 259–280, 2006.
[50] R. Kashyap, "A Bayesian comparison of different classes of dynamic models using empirical data," *IEEE Trans. Autom. Control*, vol. 22, no. 5, pp. 715–727, Oct. 1977.
[51] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
[52] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. New York, NY, USA: Academic Press, 1979.
[53] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini, "Cluster-based analysis of fMRI data," *NeuroImage*, vol. 33, no. 2, pp. 599–608, Nov. 2006.
[54] R. W. Cox, "AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages," *Comput. Biomed. Res.*, vol. 29, no. 3, pp. 162–173, 1996.
[55] R. W. Cox and J. S. Hyde, "Software tools for analysis and visualization of fMRI data," *NMR Biomed.*, vol. 10, no. 4/5, pp. 171–178, 1997.
[56] C. R. Genovese, N. A. Lazar, and T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *Neuroimage*, vol. 15, pp. 870–878, 2002.
[57] Y. Benjamini and R. Heller, "False discovery rates for spatial signals," *J. Amer. Statist. Assoc.*, vol. 102, no. 480, pp. 1272–1281, 2007.
[58] M. Smith and L. Fahrmeir, "Spatial Bayesian variable selection with application to functional magnetic resonance imaging," *J. Amer. Statist. Assoc.*, vol. 102, no. 478, pp. 417–431, 2007.
[59] K. Tabelow, J. Polzehl, H. U. Voss, and V. Spokoiny, "Analyzing fMRIexperiments with structural adaptive smoothing procedures," *NeuroImage*, vol. 33, no. 1, pp. 55–62, 2006.
[60] K. Tabelow, J. Polzehl, H. U. Voss, and V. Spokoiny, "Analyzing fMRIexperiments with structural adaptive smoothing procedures," *NeuroImage*, vol. 33, no. 1, pp. 55–62, 2006.
[61] J. Polzehl, H. U. Voss, and K. Tabelow, "Structural adaptive segmentation for statistical parametric mapping," *NeuroImage*, vol. 52, no. 2, pp. 515–523, 2010.
[62] I. A. Almodóvar-Rivera and R. Maitra, "FAST adaptive smoothed thresholding for improved activation detection in low-signal fMRI," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2821–2828, Dec. 2019.
[63] W. Cheng *et al.*, "Functional connectivity of the precuneus in unmedicated patients with depression," *Biol. Psychiatry, Cogn. Neurosci. Neuroimaging*, vol. 3, no. 12, pp. 1040–1049, Dec. 2018.
[64] Y. Jacob *et al.*, "Neural correlates of rumination in major depressive disorder: A brain network analysis," *NeuroImage: Clin.*, vol. 25, Jan. 2020, Art. no. 102142.
[65] H.-X. Zhou *et al.*, "Rumination and the default mode network: Meta-analysis of brain imaging studies and implications for depression," *NeuroImage*, vol. 206, Feb. 2020, Art. no. 116287.
[66] A. Bechara, H. Damasio, and A. R. Damasio, "Emotion, decision making and the orbitofrontal cortex," *Cereb. Cortex*, vol. 10, no. 3, pp. 295–307, Mar. 2000.
[67] F. Jollant *et al.*, "Orbitofrontal cortex response to angry faces in men with histories of suicide attempts," *Amer. J. Psychiatry*, vol. 165, no. 6, pp. 740–748, Jun. 2008.
[68] A. B. Satpute and K. A. Lindquist, "The default mode network's role in discrete emotion," *Trends Cogn. Sci.*, vol. 23, no. 10, pp. 851–864, Oct. 2019.
[69] S. R. Simon, M. Meunier, L. Piettre, A. M. Berardi, C. M. Segebarth, and D. Boussaoud, "Spatial attention and memory versus motor preparation: Premotor cortex involvement as revealed by fMRI," *J. Neurophysiol.*, vol. 88, no. 4, pp. 2047–2057, Oct. 2002.
[70] M. Koenigs, A. K. Barbey, B. R. Postle, and J. Grafman, "Superior parietal cortex is critical for the manipulation of information in working memory," *J. Neurosci.*, vol. 29, no. 47, pp. 14 980–14 986, Nov. 2009.
[71] J. J. Maller, R. H. S. Thomson, J. V. Rosenfeld, R. Anderson, Z. J. Daskalakis, and P. B. Fitzgerald, "Occipital bending in depression," *Brain*, vol. 137, no. 6, pp. 1830–1837, Jun. 2014.
[72] R. Smith *et al.*, "The role of medial prefrontal cortex in the working memory maintenance of one's own emotional responses," *Sci. Rep.*, vol. 8, no. 1, Feb. 2018, Art. no. 3460.
[73] E. T. Rolls, "The cingulate cortex and limbic systems for action, emotion, and memory," *Handbook Clin. Neurol.*, vol. 166, pp. 23–37, 2019.
[74] E. J. Bubb, L. Kinnavane, and J. P. Aggleton, "Hippocampal – Diencephalic – Cingulate networks for memory and emotion: An anatomical guide," *Brain Neurosci. Adv.*, vol. 1, 2017, Art. no. 2398212817723443.
[75] S. R. Heilbronner, B. Y. Hayden, and M. L. Platt, "Decision salience signals in posterior cingulate cortex," *Front. Neurosci.*, vol. 5, Apr. 2011, Art. no. 55.
[76] Q. Zhao, G. Zhou, S. Xie, L. Zhang, and A. Cichocki, "Tensor ring decomposition," *CoRR*, vol. abs/1606.05535, 2016.
[77] D. Zaitchik, C. Walker, S. Miller, P. LaViolette, E. Feczko, and B. C. Dickerson, "Mental state attribution and the temporoparietal junction: An fMRI study comparing belief, emotion, and perception," *Neuropsychologia*, vol. 48, no. 9, pp. 2528–2536, Jul. 2010.
[78] G. Lettieri *et al.*, "Emotionotopy in the human right temporo-parietal cortex," *Nature Commun.*, vol. 10, no. 1, Dec. 2019, Art. no. 5568.
[79] L. G. Ungerleider, S. M. Courtney, and J. V. Haxby, "A neural system for human visual working memory," *Proc. Nat. Acad. Sci.* USA, vol. 95, no. 3, pp. 883–890, Feb. 1998.

Carlos Llosa-Vite received the Bachelor of Science degree in mathematics with a minor in chemistry from The University of Arizona, Tucson, Arizona, in 2015, and the Master of Science degree in statistics from the Department of Statistics, Iowa State University, Ames, Iowa, in 2018. He is currently a PhD Candidate in statistics at Iowa State University, Ames, Iowa, where he has received the Vince Sposito Statistical Computing Award, Teaching Excellence Award, and the 2021 Three-Minute-Thesis (3MT) Competition People's Choice Award with Iowa State University, Ames, Iowa. He also received the first place winner on the 2022 Joint Statistical Meetings (JSM) Section on Imaging Statistics Student Paper Competition and he was selected as runner-up in the Statistical Methods in Imaging 2022 student paper competition. His research interests include tensor-response regression, tensor-response mixed effects models, elliptically random tensors, statistical reliability, and the statistical analysis of forensic fracture-mechanics and brain imaging data.

Ranjan Maitra received the Bachelor of Statistics (with Honours) and Master of Statistics degrees from the Indian Statistical Institute (ISI), Calcutta, India, in 1990 and 1992, respectively, and the PhD degree in statistics from the University of Washington, Seattle, Washington, in 1996. He is Professor and Associate Chair for Research in the Department of Statistics, Iowa State University, Ames, Iowa, whose faculty he joined as tenured Associate Professor, in 2003. His prior appointments were as Research Scientist with the Statistical Analysis and Data Mining Research Group, Bell Communications Research, and as tenure-track Assistant Professor and tenured Associate Professor with the Department of Mathematics and Statistics, University of Maryland, Baltimore County. His research interests include the analysis of massive datasets, computational statistics, cluster analysis, data mining, efficient simulation algorithms, image analysis and statistical computing. He received best paper and poster awards in 1995, 1996, 1996 and 1998, and the National Science Foundation's (NSF) CAREER Award, in 2003. He was elected fellow of the American Statistical Association (ASA) and elected member of the International Statistical Institute, both, in 2011. He has been editor of the multi-society-sponsored journal *Statistics Surveys* since 2010 and was the 2018–2020 area editor (Applications) for the ASA-sponsored *Statistical Analysis and Data Mining – The ASA Data Science Journal*. He was a standing committee member of the National Institutes of Health study section on the Neural Basis of Psychopathology, Addictions and Sleep Disorders from 2018–2022. His research has been funded by the NSF, National Institutes of Health, the National Institute of Justice and the National Oceanographic ad Atmospheric Administration. His mentored students have won eight ASA student paper competition awards in 2011, 2012, 2014, 2018, 2020, 2021, and 2022.