Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Individualized Multilayer Tensor Learning With an Application in Imaging Analysis

Xiwei Tang, Xuan Bi & Annie Qu

To cite this article: Xiwei Tang, Xuan Bi & Annie Qu (2020) Individualized Multilayer Tensor Learning With an Application in Imaging Analysis, Journal of the American Statistical Association, 115:530, 836-851, DOI: 10.1080/01621459.2019.1585254

To link to this article: https://doi.org/10.1080/01621459.2019.1585254

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 530, 836–851: Theory and Methods
https://doi.org/10.1080/01621459.2019.1585254

Individualized Multilayer Tensor Learning With an Application in Imaging Analysis

Xiwei Tang[a], Xuan Bi[b], and Annie Qu[c]

[a]Department of Statistics, University of Virginia, Charlottesville, VA; [b]Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, MN; [c]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL

ABSTRACT

This work is motivated by multimodality breast cancer imaging data, which is quite challenging in that the signals of discrete tumor-associated microvesicles are randomly distributed with heterogeneous patterns. This imposes a significant challenge for conventional imaging regression and dimension reduction models assuming a homogeneous feature structure. We develop an innovative multilayer tensor learning method to incorporate heterogeneity to a higher-order tensor decomposition and predict disease status effectively through utilizing subject-wise imaging features and multimodality information. Specifically, we construct a multilayer decomposition which leverages an individualized imaging layer in addition to a modality-specific tensor structure. One major advantage of our approach is that we are able to efficiently capture the heterogeneous spatial features of signals that are not characterized by a population structure as well as integrating multimodality information simultaneously. To achieve scalable computing, we develop a new bi-level block improvement algorithm. In theory, we investigate both the algorithm convergence property, tensor signal recovery error bound and asymptotic consistency for prediction model estimation. We also apply the proposed method for simulated and human breast cancer imaging data. Numerical results demonstrate that the proposed method outperforms other existing competing methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2017
Accepted February 2019

KEYWORDS
Cancer imaging; Dimension reduction; Heterogeneous modeling; High-order tensor decomposition; Multimodality integration; Spatial information.

1. Introduction

In recent years imaging analysis has encountered explosive growth due to high demand and applications in biomedical imaging for diagnosing disease status and assessing treatment outcomes. The biomedical applications of imaging analyses are especially powerful in cancer radiotherapy and neuroimaging (Fass 2008; Martino et al. 2008; Caffo et al. 2010; Zhou, Li, and Zhu 2013; Li et al. 2015). In addition, new advanced technologies in imaging analysis can be used to better understand the structures of the human body associated with biological, psychological, and clinical traits (Lindquist 2008; Lazar 2008; Friston 2009; Hinrichs et al. 2009; Kang et al. 2012), so cancer and chronic diseases can be diagnosed earlier and intervention treatments can be implemented.

This article is motivated by using multimodality optical imaging data for diagnosing early stage breast cancer prior to tumor formation. Figure 1 illustrates four-modality optical imaging data for cancerous and normal rat breast tissues, showing that a large number of tumor-associated microvesicles (TMVs; circled in red) appear spatially aligned in the cancerous tissue. One unique aspect of cancer imaging is that individuals might have imaging at various locations, and TMVs as an important biomarker are also randomly located leading to heterogeneous signal patterns. This is quite different from the brain imaging where the same regions of the brain usually share similar characteristics and functions over different subjects. Moreover, TMVs have relatively weak signals in terms of smaller size and pixel contrast values compared to modality-specific background and noise. Both great heterogeneity and weak strength make the signals difficult to be captured by a conventional population model which mainly characterizes the population-wise variation. In addition, it is also crucial to integrate information from multiple imaging modalities effectively.

In contrast to standard vector data, imaging data are usually in the form of a multidimensional array (also known as a multiorder tensor), which preserves higher-order structures containing rich spatial information. Traditional statistical models mostly treat covariates as vectors. However, fitting a model with the vectorized imaging array becomes infeasible when the array size is large. For instance, a four-modality breast cancer imaging size of 3600×3600 for each modality implicitly requires $4 \times 3600^2 = 51{,}840{,}000$ regression parameters. This leads to an ultrahigh-dimensional problem which could be unestimable even with additional regularization techniques. Most importantly, vectorization of an array is not capable of preserving the important spatial structure of imaging. Under the vectorization framework, Bayesian variable selection approaches have been developed for high-dimensional imaging regression models to identify important regions, by applying Markov random field priors to account for the spatial correlation between voxels (Penny, Trujillo-Barreto, and Friston 2005; Bowman 2007; Bowman et al. 2008; Derado, Bowman, and Kilts 2010; Li et al. 2015). However, this would rapidly increase the complexity of the prior

CONTACT Annie Qu anniequ@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

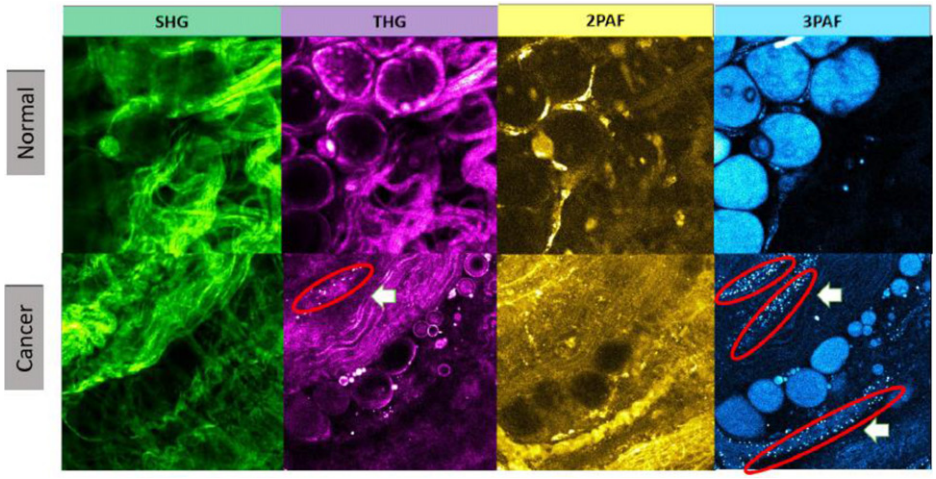

Figure 1. The four-modality microscope images for a normal rat's tissue and a cancerous rat's tissue. The increased amount and appearance of TMVs are clearly evident in the cancerous breast tissue, compared to the normal breast tissue (see red circles of TMVs).

when the tensor order increases, and could be computationally infeasible if the tensor order is high.

Alternatively, functional data analysis can be adopted to construct a two-dimensional image predictor (Reiss and Ogden 2010) in a functional regression model. Nevertheless, the extension to three dimensions or beyond could be impractical due to high-dimensional parameters arising from higher-order imaging data (Zhou, Li, and Zhu 2013). Recent developments in imaging analyses employ a two-stage strategy, that is, perform a dimension reduction such as principle component analysis (PCA) first, and then fit a regression model based on the extracted principal components (Caffo et al. 2010), however, which still requires conversion of the image into a vector first.

To use the unique higher-order structure of the image covariates, Zhou, Li, and Zhu (2013) and Li et al. (2013) proposed a tensor regression model, where the coefficients associated with imaging voxels are formulated as a tensor and assumed to be low-rank. But this potentially requires voxel-wise registration over different individuals, which is not applicable for cancer imaging. In addition, the tensor decomposition technique, such as the CANDECOMP/PARAFAC (CP) method (Beckmann and Smith 2005), is applied for processing signals from the image covariates directly (Cong et al. 2012; Mørup et al. 2006; Karahan et al. 2015; Lee et al. 2007; Miranda, Zhu, and Ibrahim 2018). Due to its population-wise low-rank constraint, it is not capable of capturing heterogeneous imaging information.

Another prevalent tool in image recognition and classification is the deep learning method, in particular the convolutional neural network (CNN, Ciresan et al. 2011; Krizhevsky, Sutskever, and Hinton 2012). In general, the CNN is powerful for image classification, which use multiple hidden layers with a variety of convolution, pooling, and activation methods. However, due to its complex architecture and the large number of parameters involved, the CNN usually requires massive training data to guarantee good performance (Keshari et al. 2018; Wagner et al. 2013; Abbasi-Asl and Yu 2017a, 2017b), especially when the imaging signals are heterogeneous and relatively weak. In our motivated cancer imaging study, such a large sample size for homogeneous samples is infeasible.

In this article, we propose an individualized multilayer tensor learning (IMTL) using additional individualized layers to incorporate heterogeneity to a higher-order tensor decomposition, which captures extra variation of the imaging covariates that cannot be characterized by a population structure. Furthermore, we extend it to multimodality data to integrate individual-specific information over different modalities and incorporate modality-specific features through different layers. We develop a scalable and efficient algorithm which can be implemented with parallel computing. In addition, we establish theoretical results regarding signal tensor recovery, identifiability, and predictive model consistency as well as the algorithm convergence property.

From a feature extraction perspective, in a sense, the decomposed basis tensor layers in the proposed approach can be viewed as some full-size feature extraction filters analogues to those in the CNN. In contrast to the CNN which applies homogeneous filters over all subjects, the individualized layers in the proposed method serve as subject-specific filters to capture additional individual-wise characteristics and heterogeneity information, which enables us to effectively capture heterogeneous outcome-associated signals such as the TMVs in breast cancer imaging data. Wang, Zhang, and Dunson (2017) proposed a similar idea for a multiple graph factorization model via decomposing the edge probability matrix to a population-shared baseline and an individual deviation. However, Wang, Zhang, and Dunson's (2017) model is not applicable to the problem considered in this article, as their approach focuses on binary data and assumes a symmetric matrix structure, while we have a multimodal three-way tensor covariate with continuous measurements of pixels.

Another advantage of the proposed method is that the individualized layer efficiently integrates multimodality images from the same subject, as it extracts a common structure over different modalities while accounting for modality-specific background noise. Multimodal imaging analysis has drawn great attention in recent years, for example, in cross-modality registration (Maes et al. 1997; Luan et al. 2008) and multiple image fusion (Wang, Sun, and Guan 2006). To the

best of our knowledge, most existing methods model the inter-modality information either after selecting features from each imaging modality separately (Zhang and Shen 2012; Liu et al. 2014), or by combining all modalities into a joint model but assuming independence between modalities (Hinrichs et al. 2011; Yuan et al. 2012). In contrast, through the individualized layers in tensor decomposition, the proposed model integrates multimodality information in both feature extraction and prediction modeling more effectively.

The article is organized as follows. Section 2 introduces the main methodology and the background. Section 3 presents the estimation and computation framework. Section 4 establishes theoretical results. Section 5 provides simulation studies. Section 6 illustrates an application for human breast cancer optical imaging data. The last section provides concluding remarks and discussion.

2. Methodology

2.1. Notation and Operations

Tensor, which also refers to a multidimensional array, plays a central role in the proposed approach. We start with providing a brief summary of tensor notation. Extensive references can be found in the review reports (Kolda 2006; Kolda and Bader 2009). A Dth-order (or D-way) tensor is a D-dimensional array $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$, where the dimensions of a tensor are also known as modes. We denote p_d $(1 \le d \le D)$ as the marginal dimension of the dth mode. In this article, we use the terms way, mode, and order interchangeably. The mode-d matricization of $\boldsymbol{\mathcal{X}}$, denoted as $\boldsymbol{\mathcal{X}}_{(d)}$, is defined as a $(p_d \times \prod_{d' \neq d} p_{d'})$-dimensional matrix, where the $(i_1, \ldots, i_D)$th element of $\boldsymbol{\mathcal{X}}$ maps to the (i_d, j)th element of $\boldsymbol{\mathcal{X}}_{(d)}$ and $j = 1 + \sum_{d' \neq d} (i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$ (Kolda and Bader 2009). Moreover, we denote vec$(\cdot)$ as a vectorizing operation which converts a tensor to a vector, where the element $x_{i_1, \ldots, i_D}$ in a D-way tensor $\boldsymbol{\mathcal{X}}$ is turned to be the $\left(i_1 + \sum_{d=2}^{D} [(i_d - 1) \prod_{j=1}^{d-1} p_j]\right)$th element in the long vector vec$(\boldsymbol{\mathcal{X}})$.

Next we introduce some useful matrix and tensor operations. Given two matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times q}$, the Kronecker product is an $mp \times nq$ matrix denoted as $\boldsymbol{A} \otimes \boldsymbol{B}$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of columns $n = q$, then the Khatri-Rao product (Rao and Mitra 1971) is defined as an $mp \times n$ matrix by $\boldsymbol{A} \odot \boldsymbol{B} = [\boldsymbol{A}_{\cdot 1} \otimes \boldsymbol{B}_{\cdot 1} \ \cdots \ \boldsymbol{A}_{\cdot n} \otimes \boldsymbol{B}_{\cdot n}]$, where $\boldsymbol{A}_{\cdot k}$ is the kth column of $\boldsymbol{A}$. The inner product $\langle \cdot, \cdot \rangle$ of two tensors with the same dimension is defined as $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \langle \text{vec}(\boldsymbol{\mathcal{A}}), \text{vec}(\boldsymbol{\mathcal{B}}) \rangle$, which follows that $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}} \rangle = \| \boldsymbol{\mathcal{A}} \|_F^2$, where $\| \cdot \|_F$ is the Frobenius norm. Moreover, an outer product "$\circ$" operating on multiple vectors $\boldsymbol{b}^1 \in \mathbb{R}^{p_1}, \ldots, \boldsymbol{b}^D \in \mathbb{R}^{p_D}$ creates a rank-1 D-way tensor $\boldsymbol{\mathcal{B}} = \boldsymbol{b}^1 \circ \boldsymbol{b}^2 \circ \ldots \circ \boldsymbol{b}^D$, where the $(i_1, i_2, \ldots, i_D)$th element of $\boldsymbol{\mathcal{B}}$ is defined as $b_{i_1, \ldots, i_D} = b^1_{i_1} b^2_{i_2} \ldots b^D_{i_D}$, and $\boldsymbol{b}^d = (b^d_1, \ldots, b^d_{p_d})'$ is the factor vector at mode d.

Consequently, a D-way tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ is of rank R if it can be represented as $\boldsymbol{\mathcal{B}} = \sum_{r=1}^{R} \boldsymbol{b}^1_r \circ \boldsymbol{b}^2_r \circ \cdots \circ \boldsymbol{b}^D_r$, where $\boldsymbol{b}^d_r$'s $(r = 1, \ldots, R)$ are p_d-dimensional vectors $(d = 1, \ldots, D)$. For ease of notation, we introduce the Kruskal operator (Kolda and Bader 2009) for a rank-R D-way tensor as $\boldsymbol{\mathcal{B}} = \boldsymbol{B}^1 \circ \boldsymbol{B}^2 \circ \cdots \circ \boldsymbol{B}^D = [\![\boldsymbol{B}^1, \boldsymbol{B}^2, \ldots, \boldsymbol{B}^D]\!]$, where $\boldsymbol{B}^d = [\boldsymbol{b}^d_1, \boldsymbol{b}^d_2, \ldots, \boldsymbol{b}^d_R] \in \mathbb{R}^{p_d \times R}$ denotes the factor matrix at mode d. This tensor rank decomposition is also known as CANDECOMP/PARAFAC (CP) decomposition, or Kruskal decomposition (Coppi and Bolasco 1988). An alternative tensor decomposition is high-order singular value decomposition, also called Tucker decomposition (Tucker 1966), which decomposes a tensor into a D-way core tensor associated with D orthonormal bases-matrices. However, the core tensor in Tucker decomposition is not guaranteed to be diagonal, and thus the tensor rank is not very clear. Therefore, we adopt the CP decomposition here.

2.2. Basic Framework and Background

The sample imaging data is usually a mixture of true signal structure and random noises. In general, we assume that

$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{\Theta}_i + \boldsymbol{\mathcal{N}}_i, \quad i = 1, \ldots, N, \tag{1}$$

where $\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is the observed D-way sample tensor covariate for the ith subject, $\boldsymbol{\Theta}_i$ is the true signal tensor associated with the outcome response, and $\boldsymbol{\mathcal{N}}_i$ is the corresponding dimensional noise background. The elements of $\boldsymbol{\mathcal{N}}_i$ are independent and identically distributed with mean of zero and variance of σ^2.

Let $f(\cdot)$ denote an appropriate feature extraction or transformation mapping on the tensor predictor $\boldsymbol{\Theta}_i \longmapsto f(\boldsymbol{\Theta}_i) : \mathbb{R}^{p_1 \times \cdots \times p_D} \longrightarrow \mathbb{R}^q$. In a sample version with recovered signal $\hat{\boldsymbol{\Theta}}_i$, let $f(\boldsymbol{\mathcal{X}}_i) = f(\hat{\boldsymbol{\Theta}}_i)$ denote the extracted sample feature. A general supervised learning stage to link the image tensor covariate $\boldsymbol{\mathcal{X}}_i$ and the observed outcome response y_i can be modeled as

$$\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f(\boldsymbol{\mathcal{X}}_i), \boldsymbol{\beta}\right), \tag{2}$$

where $\mathcal{L}(\cdot)$ is a selected loss function related to a predictive model based on the response types. For example, an L_2 loss, $\mathcal{L}(y_i, f(\boldsymbol{\mathcal{X}}_i), \boldsymbol{\beta}) = \|y_i - f(\boldsymbol{\mathcal{X}}_i)^T \boldsymbol{\beta}\|^2$, is usually applied for a linear regression model with a continuous response variable, while for the binary response in real data, we can employ a logistic regression where the corresponding $\mathcal{L}(\cdot)$ refers to the negative log-likelihood function. Alternatively, it is also flexible to utilize a machine learning model such as the support vector machine (SVM, Vapnik, Guyon, and Hastie 1995) with a hinge loss for a classification problem. Furthermore, regularization and penalty can be also incorporated to the model in (2).

Apparently, that feature extraction and data transformation is the key part to use the imaging information. One naive transformation method is to directly unfold the imaging tensor $\boldsymbol{\mathcal{X}}_i$ to a long vector via $f(\boldsymbol{\mathcal{X}}_i) = \text{vec}(\boldsymbol{\mathcal{X}}_i)$. However, the number of unknown parameters in (2) using the vectorizing covariates is $\prod_{d=1}^{D} p_d$, which is ultrahigh-dimensional and would lead to an unestimable model even applying common regularization methods. In addition, some measurement errors on the imaging data are also incorporated to the model in (2). A natural solution to solve these problems is to employ a dimension reduction technique to recover true signal structure and thus extract important features from the tensor covariates, and then fit the model in (2)

based on the extracted information. Intuitively, we consider a low-rank approximation for tensor $\boldsymbol{\mathcal{X}}_i$ as

$$\boldsymbol{\mathcal{X}}_i \approx \sum_{r=1}^{R} w_{ir}\boldsymbol{\mathcal{B}}_r, \quad i = 1, \ldots, N,$$

where $\boldsymbol{\mathcal{B}}_r$'s ($r = 1, \ldots, R$) are normalized D-way tensor bases shared by populations. This rank-R structure can be estimated by minimizing the difference between the observed imaging and approximated values under the Frobenius norm: $\sum_{i=1}^{N} \| \boldsymbol{\mathcal{X}}_i - \sum_{r=1}^{R} w_{ir}\boldsymbol{\mathcal{B}}_r \|_F^2$, and thus let $f(\boldsymbol{\mathcal{X}}_i) = (\hat{w}_{i1}, \ldots, \hat{w}_{iR})'$ be the extracted features, where the $\hat{w}_{ir}$'s are the projection loadings of the $\boldsymbol{\mathcal{X}}_i$ onto the estimated bases. However, without specifying any structures on tensor bases, the general approximation above is still likely to be unestimable and inefficient, as it is essentially a singular value decomposition problem corresponding to an $N \times \prod_{d=1}^{D} p_d$-dimensional matrix for the vectorized data.

2.3. Individualized Multilayer Tensor Decomposition

In this subsection, we propose a novel individualized multilayer tensor decomposition framework for true signal recovery and feature extraction. Following the model in (1), we have a D-way tensor covariate $\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ for each individual ($i = 1, \ldots, N$), and let $\boldsymbol{\mathcal{X}} = [\boldsymbol{\mathcal{X}}_1 \ \ldots \ \boldsymbol{\mathcal{X}}_N] \in \mathbb{R}^{N \times p_1 \times \cdots \times p_D}$ denote the aligned higher-order tensor covariate, $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1 \ \ldots \ \boldsymbol{\Theta}_N]$ denote the aligned true signal tensor, and $\boldsymbol{\mathcal{N}} = [\boldsymbol{\mathcal{N}}_1 \ \ldots \ \boldsymbol{\mathcal{N}}_N]$ denote the corresponding dimensional noise tensor. In general, the observed tensor covariate is mixed with a true signal structure and some random noises: $\boldsymbol{\mathcal{X}} = \boldsymbol{\Theta} + \boldsymbol{\mathcal{N}}$.

To motivate our model, we first start with considering a low-rank structure modeling of the grand signal tensor $\boldsymbol{\Theta}$ following a CP decomposition

$$\boldsymbol{\Theta} = \boldsymbol{W} \circ \boldsymbol{B}^1 \circ \cdots \circ \boldsymbol{B}^D = \sum_{r=1}^{R} \boldsymbol{W}_{\cdot r} \circ \boldsymbol{B}^1_{\cdot r} \circ \cdots \circ \boldsymbol{B}^D_{\cdot r}, \tag{3}$$

where $\boldsymbol{B}^d_{\cdot r}$ ($r = 1, \ldots, R$) denotes the rth column vector of mode-d factor matrix $\boldsymbol{B}^d \in \mathbb{R}^{p_d \times R}$ ($d = 1, \ldots, D$), and $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N]' \in \mathbb{R}^{N \times R}$ is the factor matrix corresponding to the dimension of individuals, and $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iR})'$ is the loading vector associated to the ith individual. To avoid the indeterminacy from scaling, we let $\bar{\boldsymbol{B}}^d_{\cdot r} = \frac{1}{\|\boldsymbol{B}^d_{\cdot r}\|}\boldsymbol{B}^d_{\cdot r}$ be the normalized factor vector at mode d. We further let $\boldsymbol{\mathcal{B}}_r = \bar{\boldsymbol{B}}^1_{\cdot r} \circ \bar{\boldsymbol{B}}^2_{\cdot r} \circ \cdots \circ \bar{\boldsymbol{B}}^D_{\cdot r}$ and $\bar{w}_{ir} = w_{ir} \prod_{d=1}^{D} \|\boldsymbol{B}^d_{\cdot r}\|$, and thus $\|\boldsymbol{\mathcal{B}}_r\|_F = 1$ (see supplementary materials A.1 for details). Then each individual tensor can be represented as

$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{\Theta}_i + \boldsymbol{\mathcal{N}}_i = \sum_{r=1}^{R} \bar{w}_{ir}\boldsymbol{\mathcal{B}}_r + \boldsymbol{\mathcal{N}}_i, \tag{4}$$

where $\boldsymbol{\mathcal{B}}_r$'s serve as population-wise rank-1 basis tensors, and thus the individual-specific feature based on the signal tensor $\boldsymbol{\Theta}_i$ is pooled as $f(\boldsymbol{\Theta}_i) = \bar{\boldsymbol{w}}_i = (\bar{w}_{i1}, \ldots, \bar{w}_{iR})'$. In the sample version, the extracted feature based on the observed sample covariates is $f(\boldsymbol{\mathcal{X}}_i) = f(\hat{\boldsymbol{\Theta}}_i) = \hat{\bar{\boldsymbol{w}}}_i$, where the estimated signals $\hat{\boldsymbol{\Theta}}_i$'s as well as the $\hat{\bar{\boldsymbol{w}}}_i$'s are obtained by minimizing an L_2-loss: $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\Theta}\|_F^2$ under the model in (3). We refer this model to the higher-order CP decomposition (HOCPD) model. The HOCPD method is powerful for reducing the tensor covariates' dimensionality, as the feature dimension is effectively reduced from $\prod_{d=1}^{D} p_d$ to R for the supervised learning in (2); however, it highly depends on the low-rank assumption on the grand signal tensor, which could fail to capture complex tensor data information if there is significant heterogeneous variation arising from unique individuals. Moreover, it is infeasible to capture heterogeneity simply by increasing the rank in an HOCPD model, as it might lead to nonidentifiability and unstable model estimation.

In the following, we propose an individualized multilayer tensor learning (IMTL) method to incorporate heterogeneous structures to tensor decomposition. For the ith individual, we assume

$$\boldsymbol{\Theta}_i = \sum_{r=1}^{R} \bar{w}_{ir}\boldsymbol{\mathcal{B}}_r + \boldsymbol{\mathcal{S}}_i, \quad \text{s.t.} \quad \langle \boldsymbol{\mathcal{B}}_r, \boldsymbol{\mathcal{S}}_i \rangle = 0, \ 1 \le r \le R, \tag{5}$$

where $\boldsymbol{\mathcal{B}}_r$'s are the population-shared rank-1 bases analogous to those in (4), and $\boldsymbol{\mathcal{S}}_i = \boldsymbol{s}_i^1 \circ \boldsymbol{s}_i^2 \circ \cdots \circ \boldsymbol{s}_i^D$ is an individualized rank-1 structure. With normalized vectors $\bar{\boldsymbol{s}}_i^d$'s and $u_i = \prod_{d=1}^{D} \|\boldsymbol{s}_i^d\|$, we can rewrite $\boldsymbol{\mathcal{S}}_i = u_i\bar{\boldsymbol{\mathcal{S}}}_i$ where $\bar{\boldsymbol{\mathcal{S}}}_i = \bar{\boldsymbol{s}}_i^1 \circ \bar{\boldsymbol{s}}_i^2 \circ \cdots \circ \bar{\boldsymbol{s}}_i^D$. Given observed sample tensor covariates, the latent parameters in model (5) are estimated through

$$\min_{\boldsymbol{w}_i, \boldsymbol{\mathcal{B}}_r, \mu_i, \bar{\boldsymbol{\mathcal{S}}}_i} \sum_{i=1}^{N} \|\boldsymbol{\mathcal{X}}_i - \sum_{r=1}^{R} \bar{w}_{ir}\boldsymbol{\mathcal{B}}_r - \mu_i\bar{\boldsymbol{\mathcal{S}}}_i\|_F^2,$$
$$\text{s.t.} \quad \langle \boldsymbol{\mathcal{B}}_r, \bar{\boldsymbol{\mathcal{S}}}_i \rangle = 0, \ 1 \le r \le R.$$

The orthogonal constraint is imposed to ensure the identifiability between the homogeneous layers and the heterogeneous layers, and also to improve the computational stability. Based on the recovered signal $\hat{\boldsymbol{\Theta}}_i$, the extracted individual feature is $f(\hat{\boldsymbol{\Theta}}_i) = (\hat{\bar{\boldsymbol{w}}}_i', \hat{u}_i, \{\hat{\bar{\boldsymbol{s}}}_i'^d\}_{d=1}^{D})'$, containing the weights of the population-shared layers $\hat{\bar{\boldsymbol{w}}}_i = (\bar{w}_{i1}, \ldots, \bar{w}_{iR})'$, the weight of the individualized layer $\hat{\mu}_i$, and the decomposed factors of the individualized layers $\hat{\bar{\boldsymbol{s}}}_i^d$'s, respectively.

In contrast to the HOCPD in (4), each signal tensor $\boldsymbol{\Theta}_i$ in (5) is characterized by two different layers, one consisting of a linear combination of homogeneous structures ($\boldsymbol{\mathcal{B}}_r$'s), and another containing an individualized structure ($\boldsymbol{\mathcal{S}}_i$) capturing the heterogeneity of individual features. Intuitively, the individualized layer $\boldsymbol{\mathcal{S}}_i$ is the "best" rank-1 structure extracted from the "residual" tensor of a higher-order decomposition. This additional layer enables us to capture the individual-specific information of the tensor covariate which is not characterized by a population structure $\sum_{r=1}^{R} \bar{w}_{ir}\boldsymbol{\mathcal{B}}_r$, while the rank-1 structure in (5) avoids overfitting on the individualized layers.

Additionally, in practice, the informative heterogeneous signals could be relatively weak compared to the population-wise background and noise. For example, in real data, the important biomarker TMVs are discrete and random spatially distributed, while the signal strength is also quite weak. Therefore, they are very likely to be washed out if only low-rank homogeneous bases are used. In some sense, the homogeneous bases in the

population layer serve as some full-size filters to remove those homogeneous variation, so the subject-specific signals which deviate from the low-rank homogeneous background can be captured. Consequently, the population layer and the individual layer effectively coordinate to extract informative features from sample tensor covariates and thus improve model prediction power.

2.4. Multilayer Tensor Reconstruction for Multimodality Data

Multimodality imaging data are widely adopted and have drawn great attentions in recent years. In our motivating application, optical imaging produces multiple images using different wavelengths of light in one examination. Although a single modality model is applicable, there is a critical need to integrate all information collected from multiple modalities, so important features associated with clinical outcomes can be extracted more effectively and the prediction power can be greatly enhanced. In the following, we develop a multimodality imaging tensor model which extends the individualized multilayer approach in (5) to incorporate different sources of modality information.

We consider an M-modality tensor covariate $\boldsymbol{\Theta} = [\boldsymbol{\Theta}^{(1)} \ \cdots \ \boldsymbol{\Theta}^{(M)}]$, where $\boldsymbol{\Theta}^{(m)} \in \mathbb{R}^{N\times p_1\times\cdots\times p_D}$ $(1 \le m \le M)$. For a single modality m, we propose a multilayer tensor decomposition

$$\boldsymbol{\Theta}^{(m)} = \boldsymbol{W} \circ \boldsymbol{B}^{(m),1} \circ \cdots \circ \boldsymbol{B}^{(m),D} + \boldsymbol{\mathcal{S}},$$

where $\boldsymbol{B}^{(m),d}$'s are modality-specific factors, and $\boldsymbol{\mathcal{S}} = [\boldsymbol{\mathcal{S}}_1 \ \boldsymbol{\mathcal{S}}_2 \ \cdots \ \boldsymbol{\mathcal{S}}_N] \in \mathbb{R}^{N\times p_1\times\cdots\times p_D}$ contains heterogeneous layers. With normalized factors, each individual tensor can be rewritten as

$$\boldsymbol{\Theta}_i^{(m)} = \sum_{r=1}^{R} \bar{w}_{ir}^{(m)} \boldsymbol{\mathcal{B}}_r^{(m)} + \mu_i \bar{\boldsymbol{\mathcal{S}}}_i, \ \text{ s.t. } \left\langle \boldsymbol{\mathcal{B}}_r^{(m)}, \bar{\boldsymbol{\mathcal{S}}}_i \right\rangle = 0, \ 1 \le r \le R, \tag{6}$$

where $\{\boldsymbol{\mathcal{B}}_r^{(m)} = \bar{\boldsymbol{B}}_{\cdot r}^{(m),1} \circ \cdots \circ \bar{\boldsymbol{B}}_{\cdot r}^{(m),D}\}_{r=1}^{R}$ is the set of bases for the mth modality, and $\boldsymbol{\mathcal{S}}_i = \mu_i \bar{\boldsymbol{\mathcal{S}}}_i$ is an individualized layer shared by different modalities $\boldsymbol{\Theta}_i^{(m)}$ $(m = 1, \ldots, M)$ of the ith subject. Similarly, the proposed model in (6) can be estimated by minimizing a sum of squared loss: $\sum_{m=1}^{M} \| \boldsymbol{\mathcal{X}}^{(m)} - \boldsymbol{\Theta}^{(m)} \|_F^2$ with the constraints. Let $\boldsymbol{\Theta}_i = [\boldsymbol{\Theta}_i^{(1)} \ldots \boldsymbol{\Theta}_i^{(M)}]$ be the multimodality tensor signal for the ith subject, and denote $\hat{\boldsymbol{\Theta}}_i$ as the sample estimator based on (6). The extracted subject-wise feature is $f(\hat{\boldsymbol{\Theta}}_i) = (\hat{\hat{\boldsymbol{w}}}_i, \hat{\mu}_i, \{\hat{\hat{\boldsymbol{s}}}_i^d\}_{d=1}^{D})'$, containing the weights of the modality specific layers $\hat{\hat{\boldsymbol{w}}}_i = \{\hat{\hat{w}}_{ir}^{(m)}\}_{1\le r\le R}^{1\le m\le M}$, the weight of the individualized layer $\hat{\mu}_i$, and the decomposed factors of the individualized layer $\hat{\hat{\boldsymbol{s}}}_i^d$'s, respectively.

For the proposed multimodality model in (6), the individualized layer $\boldsymbol{\mathcal{S}}_i$ is a common low-rank structure capturing the individual-specific spatial pattern shared by different modalities. Similarly, for a given subject, if we consider removing the modality-specific layers $\sum_{r=1}^{R} \bar{w}_{ir}^{(m)} \boldsymbol{\mathcal{B}}_r^{(m)}$ from each modality $\boldsymbol{\Theta}_i^{(m)}$ first, and then aligning the "residuals" from multiple modalities to a higher-order tensor ($\mathbf{R}^{p_1\times\cdots\times p_D\times M}$), then the individualized layer can be interpreted as a tensor slice along the modality mode of the "best" rank-1 structure extracted from the higher-order residual tensor.

In addition to characterizing the heterogeneous structures analogues to the singlemodality case, the individualized layer also makes it feasible to integrate the information of multiple modalities effectively. Although the individual-specific signals are randomly distributed across different individuals, their spatial pattern is shared by different modalities from the same individual. Therefore, the individualized layer extracted over multiple modalities aggregates the information regarding their common structure and thus captures this individual-specific feature more accurately. Figure 2 provides an illustration of the individualized layers and the modality-specific layers on the four-modality breast cancer images.

Alternatively, we can simply align the four modalities together to a higher-order tensor predictor with a size of $p_1 \times \cdots \times p_D \times M$, and then apply the proposed singlemodality model in (5). However, this is inadequate to capture the variation among different modalities. In the following, we compare the singlemodality model in (5) with the integrated tensor and multimodality model in (6). Let $\boldsymbol{\Theta}_i = [\boldsymbol{\Theta}_i^{(1)}, \ldots, \boldsymbol{\Theta}_i^{(M)}] \in \mathbf{R}^{p_1\times\cdots\times p_D\times M}$ denote the multimodality integrated tensor. With a singlemodality model, we have

$$\boldsymbol{\Theta}_i = \sum_{r=1}^{R} w_{ir} \boldsymbol{b}_r^1 \circ \cdots \circ \boldsymbol{b}_r^D \circ \boldsymbol{b}_r^{\text{modal}} + \boldsymbol{s}_i^1 \circ \cdots \circ \boldsymbol{s}_i^D \circ \boldsymbol{s}_i^{\text{modal}},$$

where $\boldsymbol{b}_r^{\text{modal}} = (b_{1r}^{\text{modal}}, \ldots, b_{Mr}^{\text{modal}})'$ and $\boldsymbol{s}_i^{\text{modal}} = (s_{i1}^{\text{modal}}, \ldots, b_{iM}^{\text{modal}})'$ are both basis vectors on the mode of modality. Hence, for the mth modality $(m = 1, 2, 3, 4)$, the corresponding tensor slice ($p_1 \times \cdots \times p_D$-dimensional tensor) can be written as

$$\boldsymbol{\Theta}_i^{(m)} = \sum_{r=1}^{R} w_{ir} b_{mr}^{\text{modal}} \boldsymbol{b}_r^1 \circ \cdots \circ \boldsymbol{b}_r^D + s_{im}^{\text{modal}} \boldsymbol{s}_i^1 \circ \cdots \circ \boldsymbol{s}_i^D.$$

This implies that the population-shared basis layers $\boldsymbol{\mathcal{B}}_r = \boldsymbol{b}_r^1 \circ \cdots \circ \boldsymbol{b}_r^D$ $(1 \le r \le R)$ are the same across all modalities up to scaling and permutation. In contrast, the multimodality model in (6) allows different modality-specific layers $\boldsymbol{\mathcal{B}}_r^{(m)}$ across different modalities to capture the modality-wise heterogeneity. In fact, the singlemodality model on the integrated data is a special case of the multimodality model (6), with additional constraints on all modalities sharing the same set of population bases.

3. Computation

In this section, we address the estimation problem of multilayer decomposition in (6) for the proposed IMTL model. In contrast to traditional CP decomposition, incorporating modality layers and individual layers of the proposed method significantly increases the computation cost, and conventional algorithms for tensor decomposition are not necessarily scalable in our situation. Therefore, we propose a bi-level block improvement algorithm which alternately updates different layers and apply a maximum block improvement (MBI) strategy for the estimation of each layer.

The proposed model in (6) yields a constrained optimization requiring orthogonality between the modality layers and the individual layers. Hence, we employ a penalization method (Nocedal and Wright 2006) to achieve orthogonality

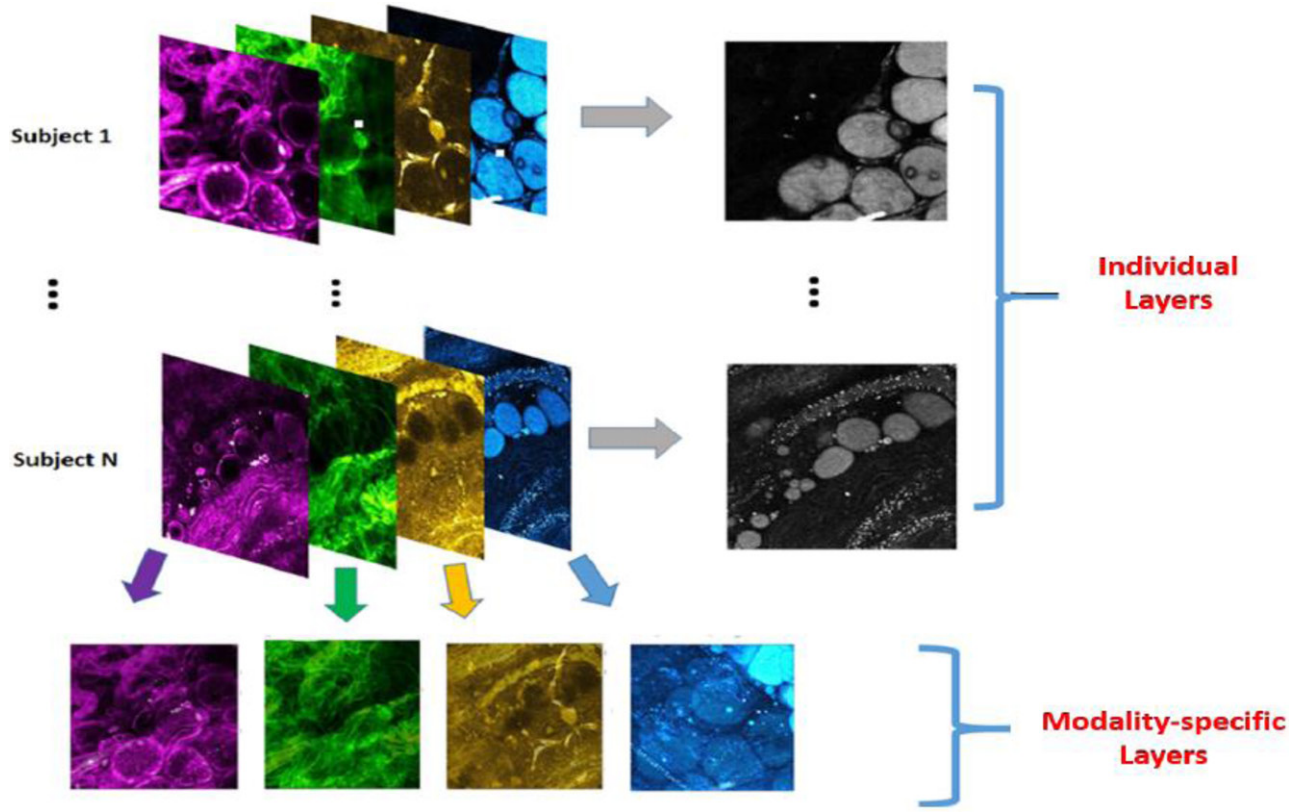

Figure 2. An illustration of the individualized layers and the modality-specific layers for four-modality optical images of the breast cancer tissues.

constrained optimization, that is, to minimize the objective function

$$L\left(\boldsymbol{W}, \{\boldsymbol{B}^{(m),d}\}_{m's,d's}, \{\boldsymbol{s}_i^d\}_{i's,d's} \mid \boldsymbol{\mathcal{X}}\right) = \sum_{m=1}^{M} \|\boldsymbol{\mathcal{X}}^{(m)} - \boldsymbol{\Theta}^{(m)}\|_F^2 + \lambda_s \sum_{m,i,r} \langle \boldsymbol{\mathcal{B}}_r^{(m)}, \boldsymbol{\mathcal{S}}_i \rangle^2. \tag{7}$$

Note that the orthogonal property of the decomposed layers is not critical here as it only serves to ensure identifiability. Although the penalization method cannot guarantee an exact orthogonality of the estimated layers, it is sufficient for pursuing the layers' identifiability in practice.

The objective function in (7) is convex and differentiable with respect to each separate block of parameters, which makes it applicable to apply a block-wise updating procedure. However, the traditional tensor decomposition algorithms, for example, alternating least squares (ALS, Carroll and Chang 1970) and block relaxation algorithm (De Leeuw 1994; Lange 2010), are nearly infeasible and unscalable due to the large number $((M + N)D + 1)$ of blocks, and updating one block at each iteration could lead to poor convergence performance. In addition, the estimation of different blocks (layers) use different parts of the data, for example, $\boldsymbol{B}^{(m),d}$'s are estimated within each modality and $\boldsymbol{s}_i^d$'s are estimated within each subject. This allows for alternately estimating different layers, and makes parallel computing feasible, which significantly improves the computational scalability. Finally, in general, for $D > 2$, the ALS-type cyclical algorithm is not guaranteed to converge to a stationary point (Chen et al. 2012). Therefore, we propose a bi-level algorithm, that is, within the estimation of each layer, we apply the MBI strategy (Chen et al. 2012) in estimating specific blocks, and then we update each layer alternately.

Specifically, given the individual-layer factors, we let $\boldsymbol{\mathcal{X}}^{(m)} = \boldsymbol{\mathcal{X}}^{(m)} - \hat{\boldsymbol{\mathcal{S}}}$ and alternately update blocks of $\boldsymbol{W}$ and $\{\boldsymbol{B}^{(m),d}\}$'s; that is,

$$\hat{\boldsymbol{W}}^{[new]} = \arg\min_{\boldsymbol{W}} \sum_{m=1}^{M} \left\| \tilde{\boldsymbol{\mathcal{X}}}_{(1)}^{(m)} - \boldsymbol{W}\left(\hat{\boldsymbol{B}}^{(m),D} \odot \cdots \odot \hat{\boldsymbol{B}}^{(m),1}\right)^T \right\|_F^2, \tag{8}$$

where $\tilde{\boldsymbol{\mathcal{X}}}_{(1)}^{(m)}$ is the mode-1 matricization. The sub-optimization turns to be a ridge-type problem and thus has a unique explicit solution. Next, given $\hat{\boldsymbol{W}}$, we update $\{\boldsymbol{B}^{(m),d}\}_{d's}$ through parallel computing across different modalities $(m = 1, \ldots, M)$, and apply the MBI strategy. Let

$$L^{(m)}\left(\boldsymbol{B}^{(m),d} \mid \{\boldsymbol{B}^{(m),d'}\}_{d' \neq d}\right) = \left\| \tilde{\boldsymbol{\mathcal{X}}}_{(d+1)}^{(m)} - \boldsymbol{B}^{(m),d}\left(\boldsymbol{B}_{-d}^{(m)} \odot \hat{\boldsymbol{W}}\right)^T \right\|_F^2 + \lambda_s \sum_i \left\langle \boldsymbol{B}^{(m),d}, \hat{\boldsymbol{\mathcal{S}}}_{i,(d)} \boldsymbol{B}_{-d}^{(m)} \right\rangle^2$$

be the reduced objective function with respect to $\boldsymbol{B}^{(m),d}$ while fixing all other blocks, where $\boldsymbol{B}_{-d}^{(m)} = \boldsymbol{B}^{(m),D} \odot \cdots \boldsymbol{B}^{(m),d+1} \odot \boldsymbol{B}^{(m),d-1} \cdots \odot \boldsymbol{B}^{(m),1}$ and $\hat{\boldsymbol{\mathcal{S}}}_{i,(d)}$ is the mode-d matricization of $\hat{\boldsymbol{\mathcal{S}}}_i$. Given the latest iterate $\{\boldsymbol{B}^{(m),1,[t]}, \ldots, \boldsymbol{B}^{(m),D,[t]}\}$, we calculate

$$\begin{aligned} \boldsymbol{B}^{*(m),d} &= \arg\min_{\boldsymbol{B}^{(m),d}} L^{(m)}(\boldsymbol{B}^{(m),d} \mid \{\boldsymbol{B}^{(m),d',[t]}\}_{d' \neq d}), \\ d_m^* &= \arg\min_{1 \leq d \leq D} L^{(m)}(\boldsymbol{B}^{*(m),d} \mid \{\boldsymbol{B}^{(m),d',[t]}\}_{d' \neq d}), \end{aligned} \tag{9}$$

and update only the block $\boldsymbol{B}^{(m),d,[t]}$ at the mode $d = d_m^*$ for maximum improvement.

Similarly, given the modality-layer factors, the individualized layers can be estimated within each subject using parallel computing. For each subject i, let $\bar{\boldsymbol{\mathcal{X}}}_i^{(m)} = \boldsymbol{\mathcal{X}}_i^{(m)} - \sum_{r=1}^{R} \hat{w}_{ir} \hat{\boldsymbol{\mathcal{B}}}_r^{(m)}$, and let

$$L_i\big(\boldsymbol{s}_i^d \mid \{\boldsymbol{s}_i^{d'}\}_{d' \neq d}\big) = \sum_{m=1}^{M} \|\bar{\boldsymbol{\mathcal{X}}}_{i,(d)}^{(m)} - \boldsymbol{s}_i^d (\boldsymbol{S}_{i,-d})^T\|_F^2 + \lambda_s \sum_{m,r} \langle \boldsymbol{s}_i^d,\ \hat{\boldsymbol{\mathcal{B}}}_{r,(d)}^{(m)} \boldsymbol{S}_{i,-d} \rangle^2$$

be the reduced objective function with respect to $\boldsymbol{s}_i^d$, where $\boldsymbol{S}_{i,-d} = \boldsymbol{s}_i^1 \odot \cdots \boldsymbol{s}_i^{d-1} \odot \boldsymbol{s}_i^{d+1} \cdots \odot \boldsymbol{s}_i^D$ and $\hat{\boldsymbol{\mathcal{B}}}_{r,(d)}^{(m)}$ is the mode-d matricization. Given the latest iterate $\{\boldsymbol{s}_i^{1,[t]}, \ldots, \boldsymbol{s}_i^{D,[t]}\}$, calculate

$$\boldsymbol{s}_i^{*d} = \arg\min_{\boldsymbol{s}_i^d} L_i(\boldsymbol{s}_i^d | \{\boldsymbol{s}_i^{d',[t]}\}_{d' \neq d}), \quad \text{and}$$
$$d_i^* = \arg\min_{1 \le d \le D} L_i(\boldsymbol{s}_i^{*d} | \{\boldsymbol{s}_i^{d',[t]}\}_{d' \neq d}), \tag{10}$$

and update only mode-d_i^* block. The detailed algorithm is summarized in following Algorithm 1.

For any objective function $f(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$ with blocks of parameters $\{\boldsymbol{A}_i\}_{i=1}^n$, a point of $(\boldsymbol{A}_1^*, \ldots, \boldsymbol{A}_n^*)$ in the parameter space is defined as a block-wise stationary point of $f(\cdot)$ if for any block, there is $\boldsymbol{A}_i^* = \arg\min_{\boldsymbol{A}_i} f(\boldsymbol{A}_1^*, \ldots, \boldsymbol{A}_{i-1}^*, \boldsymbol{A}_i, \boldsymbol{A}_{i+1}^*, \ldots, \boldsymbol{A}_n^*)$.

Let $\{\boldsymbol{W}, \boldsymbol{B}^{(m),d}, \boldsymbol{s}_i^d\}_{m,i,d}$ be an element-wise collection of all parameters and let Ω be the corresponding parameter space. We provide a global convergence property of Algorithm 1 as follows.

Lemma 1. Assume Ω is compact, then any accumulative point of the iterations from Algorithm 1, say $\{\boldsymbol{W}^*, \boldsymbol{B}^{*(m),d}, \boldsymbol{s}_i^{*d}\}_{m's,i's,d's}$, is a block-wise stationary point of the objective function in (7).

In general, multiple initializations are suggested to obtain a sound optimum in (7). The normalization on modes' factor vectors can be performed after the last iteration. Given the estimated tensor signals $\hat{\boldsymbol{\Theta}}_i$'s, for binary response, we employ an L_1-penalized logistic regression model with the extracted sample features $f(\hat{\boldsymbol{\Theta}}_i)$'s in (2) for prediction. The rank of the population-shared layers and the other tuning parameter are selected based on a grid search to minimize the prediction errors on the validation set or via a cross-validation method. More details and discussion about tuning procedure and the prediction on a new subject are provided in the supplementary materials.

4. Theory

In this section, we develop the theoretical framework for the proposed model regarding both tensor signal recovery and prediction modeling. First, we introduce conditions to ensure identifiability of the proposed multilayer tensor modeling. Furthermore, we provide an accurate error bound for the recovered signal tensor and show that the estimated signal tensor converges to the true one. Finally, we present the theoretical results of the supervised learning stage to evaluate prediction performance. All proofs are provided in supplementary materials Section A.

We first address the tensor modeling identifiability issue before establishing the statistical property, which is critical for tensor representation. In particular, we provide sufficient conditions to achieve identifiable layers in the proposed multilayer tensor model, which can be verified easily in practice. For ease of notation, the following discussion focuses on the single modality model while the presented conditions can be easily extended to the multimodality model.

In the proposed framework, potential unidentifiability could occur in the multilayer CP decomposition of tensor predictors in (5), which can be attributed to three aspects. The first two indeterminacies arise from scaling and permutation, and the last aspect is the nonuniqueness of the CP decomposition for a tensor, that is, the possibility that more than one combination of population layers and individualized layers can lead to the underlying true image tensor. In the proposed model, the scaling indeterminacy, which refers to possible rescaling over different modes' factor vectors for each layer, is eliminated by imposing a unit-norm constraint on parameterization, that is, $\|\bar{\boldsymbol{B}}_{\cdot r}^d\|_2 = 1$ and $\|\bar{\boldsymbol{s}}_i^d\|_2 = 1$. Moreover, the permutation indeterminacy refers to the arbitrary reordering of the population bases. To address this point, we could align the population bases according to a

Algorithm 1 A bi-level block improvement algorithm with parallel computing

1.(*Initialization*). Set $t = 1$ and the tuning parameter λ_s. Set initial values for $\boldsymbol{W}^{[0]}$, $\boldsymbol{B}^{(m),d,[0]}$'s ($1 \le m \le M; 1 \le d \le D$). Set $\boldsymbol{s}_i^{d,[0]} = \boldsymbol{0}$ ($1 \le i \le N; 1 \le d \le D$).

2.(*Modality layers*). At the tth iteration, given $\{\boldsymbol{s}_i^{d,[t-1]}\}'s$, estimate $\boldsymbol{W}^{[t]}$ and $\{\boldsymbol{B}^{(m),d,[t]}\}'s$.

(i) Set $\boldsymbol{W}^{[t]} \leftarrow \boldsymbol{W}^{[t-1]}$, $\boldsymbol{B}^{(m),d,[t]} \leftarrow \boldsymbol{B}^{(m),d,[t-1]}$ (m's, d's).

(ii) Fixing all $\boldsymbol{B}^{(m),d,[t]}$'s, update $\boldsymbol{W}^{[t]_{\text{new}}}$ through (8).

(iii) Fixing $\boldsymbol{W}^{[t]}$, for each modality m, (a) calculate d_m^*'s and $\boldsymbol{B}^{*(m),d}$'s based on (9); (b) assign $\boldsymbol{B}^{(m),d,[t]_{\text{new}}} \leftarrow \boldsymbol{B}^{*(m),d}$ if $d = d_m^*$, and $\boldsymbol{B}^{(m),d,[t]_{\text{new}}} \leftarrow \boldsymbol{B}^{(m),d,[t]}$ if $d \neq d_m^*$.

(iv) Stop iteration if $\frac{1}{NR}\|\boldsymbol{W}^{[t]_{\text{new}}} - \boldsymbol{W}^{[t]}\|_F^2 + \frac{1}{MR\prod_{d=1}^D p_d} \sum_{m,D} \| \times \boldsymbol{B}^{(m),d,[t]_{\text{new}}} - \boldsymbol{B}^{(m),d,[t]}\|_F^2 \le 10^{-4}$, otherwise assign $\boldsymbol{W}^{[t]} \leftarrow \boldsymbol{W}^{[t]_{\text{new}}}$, $\boldsymbol{B}^{(m),d,[t]} \leftarrow \boldsymbol{B}^{(m),d,[t]_{\text{new}}}$ (m's, d's), and go to Step 2(ii).

3. (*Individualized layers*). Given $\boldsymbol{W}^{[t]}$, $\boldsymbol{B}^{(m),d,[t]}$'s, for individual i ($1 \le i \le N$), estimate $\boldsymbol{s}_i^{d,[t]}$'s.

(i) Set $\boldsymbol{s}_i^{d,[t]} \leftarrow \boldsymbol{s}_i^{d,[t-1]}$ (d's).

(ii) (a) Calculate d_i^*'s and $\boldsymbol{s}_i^{*d}$'s based on (10); (b) assign $\boldsymbol{s}_i^{d,[t]_{\text{new}}} \leftarrow \boldsymbol{s}_i^{*d}$ if $d = d_i^*$, and $\boldsymbol{s}_i^{d,[t]_{\text{new}}} \leftarrow \boldsymbol{s}_i^{d,[t]}$ if $d \neq d_i^*$.

(iii) Stop iteration if $\frac{1}{\prod_{d=1}^D p_d} \sum_d \|\boldsymbol{s}_i^{d,[t]_{\text{new}}} - \boldsymbol{s}_i^{d,[t]}\|_F^2 \le 10^{-4}$, otherwise assign $\boldsymbol{s}_i^{d,[t]} \leftarrow \boldsymbol{s}_i^{d,[t]_{\text{new}}}$, and go to Step 3(ii).

4. (*Stopping criterion*). Stop if $\frac{\|\boldsymbol{\Theta}^{[t]} - \boldsymbol{\Theta}^{[t-1]}\|_F^2}{NM\prod_{d=1}^D p_d} \le 10^{-3}$, otherwise set $t \leftarrow t+1$ and go to Step 2.

descending order of the first element of mode-1 factor vectors, that is, $\bar{\boldsymbol{B}}_{11}^1 \geq \bar{\boldsymbol{B}}_{12}^1 \geq \cdots \geq \bar{\boldsymbol{B}}_{1R}^1$.

After controlling the scaling and the permutation, in general, the CP decomposition of a tensor could still be nonunique, due to the possibility of multiple combinations of rank-one tensors in decomposition. Although various identifiability conditions have been presented for a conventional CP decomposition (Coppi and Bolasco 1988; Sidiropoulos and Bro 2000; Kolda and Bader 2009), they are infeasible here as all of them require checking each individual tensor $\boldsymbol{\Theta}_i$ $(1 \leq i \leq N)$ separately, which is not effective in practice, especially when the sample size N is increasing.

In the following, we provide a much weaker sufficient condition based on the integrated higher-order tensor without imposing any additional constraints on parameterization. We first introduce the concept of a k-rank of a matrix according to Coppi and Bolasco (1988). Specifically, the k-rank of a matrix $\boldsymbol{A}$, denoted as $\mathcal{K}_{\boldsymbol{A}}$ is defined as $\mathcal{K}_{\boldsymbol{A}} = \max\{k : \text{any } k \text{ columns of } \boldsymbol{A} \text{ are linearly independent}\}$. Let $\boldsymbol{\Theta}_{[1:n]}$ denote an integrated $(D+1)$-way tensor combining n individual tensors. Without loss of generality, we assume $\boldsymbol{\Theta}_{[1:n]} = [\boldsymbol{\Theta}_1 \ \cdots \ \boldsymbol{\Theta}_n]$. There is a $(R+n)$-rank representation for the integrated tensor

$$\boldsymbol{\Theta}_{[1:n]} = \sum_{r=1}^{R} \boldsymbol{w}_r^{[1:n]} \circ \boldsymbol{B}_{\cdot r}^1 \circ \cdots \circ \boldsymbol{B}_{\cdot r}^D + \sum_{i=1}^{n} \tilde{\boldsymbol{u}}_i^{[1:n]} \circ \boldsymbol{s}_i^1 \circ \cdots \circ \boldsymbol{s}_i^D,$$

where $\boldsymbol{w}_r^{[1:n]} = (w_{1r}, \ldots, w_{nr})'$ and $\tilde{\boldsymbol{u}}_i^{[1:n]} = (\underbrace{0, \ldots, 0}_{1,\ldots,i-1}, \underbrace{1}_{i}, \underbrace{0, \ldots, 0}_{i+1,\ldots,n})'$. For $1 \leq d \leq D$, let $\tilde{\boldsymbol{B}}_{[1:n]}^d = [\boldsymbol{b}_1^d \ \ldots \ \boldsymbol{b}_R^d \ \boldsymbol{s}_1^d \ \ldots \ \boldsymbol{s}_n^d]$ denote the mode-d factor matrix for integrated tensor $\boldsymbol{\Theta}_{[1:n]}$. We have the following proposition providing a sufficient condition for the identifiability of the multilayer decomposition in (6).

Proposition 1. If there exist n individual tensors $(2 \leq n \leq N)$ in $\{\boldsymbol{\Theta}_i, 1 \leq i \leq N\}$, such that,

$$\sum_{d=1}^{D} \mathcal{K}_{\tilde{\boldsymbol{B}}_{[1:n]}^d} \geq 2R + n + D$$

holds for the integrated high-order tenor $\boldsymbol{\Theta}_{[1:n]}$, then the multilayer decomposition in (5) is unique given the unit-norm and the ordering constraints on the factor vectors.

In the proof of Proposition 1, we show that any rank-1 CP decomposition for a D-way $(D \geq 2)$ tensor is unique up to scaling indeterminacy, which implies the identifiability of the individualized layer given the population bases. The above condition is relatively weak compared to those in Coppi and Bolasco (1988), Sidiropoulos and Bro (2000), and Kolda and Bader (2009), and it is easy to satisfy as it applies for an arbitrary n. For example, if $n = 2$ and the factor matrices are of full rank, then the condition in Proposition 1 holds as long as $D \geq 2$.

Next we establish the theoretical properties for the recovered tensor signal $\hat{\boldsymbol{\Theta}}$ based on the observed sample tensor covariates $\boldsymbol{\mathcal{X}}_i^{(m)}$'s. We denote $\boldsymbol{\gamma} = \left(\text{vec}(\boldsymbol{W})', \{\text{vec}(\boldsymbol{B}^{(m),d})\}_{m,d}', \{\boldsymbol{s}_i^d\}_{i,d}'\right)'$ as the vector of all latent variable parameters. It is straightforward that $\dim(\boldsymbol{\gamma}) = MR(\sum_{d=1}^D p_d) + N(R + \sum_{d=1}^D p_d)$. Let $\mathcal{I} = \{\boldsymbol{\mathcal{X}}_i^{(m)}\}_{i,m}$ denote a collection of all observed D-way tensors and $|\mathcal{I}|$ be the cardinality measure of $\mathcal{I}$. That is, $|\mathcal{I}|$ denotes the number of all singlemodality images, which equals NM if there is no missing.

Furthermore, let $\boldsymbol{\Theta}^{(m)} = (\theta_{i,j_1\cdots j_D}^{(m)})$ and we have $\theta_{i,j_1\cdots j_D}^{(m)} = \sum_{r=1}^{R} w_{ir} b_{j_1 r}^{(m),1} \cdots b_{j_D r}^{(m),D} + s_{i,j_1}^1 \cdots s_{i,j_D}^D$ according to the proposed multilayer model in (6). We assume that $\mathbf{E}[X_{i,j_1\ldots j_D}^{(m)}] = \theta_{i,j_1\ldots j_D}^{(m)}$, where $X_{i,j_1\ldots j_D}^{(m)}$ denotes an element of the sample image tensor $\boldsymbol{\mathcal{X}}_i^{(m)}$, for example, a pixel in the image, and $x_{i,j_1\ldots j_D}^{(m)}$ denotes an observed value. Then, the number of all sample elements' observations is $|\mathcal{I}| \prod_{d=1}^D p_d$, which increases as the number of images increases. In the following, for the $(j_1, \ldots, j_D)$th element of image $\boldsymbol{\mathcal{X}}_i^{(m)}$, we define the loss function

$$\begin{aligned} l(\Theta | X_{i,j_1\cdots j_D}^{(m)} = x_{i,j_1\cdots j_D}^{(m)}) &= (x_{i,j_1\cdots j_D}^{(m)} - \theta_{i,j_1\cdots j_D}^{(m)})^2 \\ &= (x_{i,j_1\cdots j_D}^{(m)} - \sum_{r=1}^{R} w_{ir} b_{j_1 r}^{(m),1} \cdots b_{j_D r}^{(m),D} - s_{ij_1}^1 \cdots s_{ij_D}^D)^2. \end{aligned}$$

Consequently, we assume that the overall objective function is an additive form of the loss function and the penalty function, that is,

$$L(\Theta | \boldsymbol{\mathcal{X}}) = \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j_1=1}^{p_1} \cdots \sum_{j_D=1}^{p_D} l(\Theta | x_{i,j_1\cdots j_D}^{(m)}) + \lambda_{|\mathcal{I}|} p(\boldsymbol{\Theta}),$$

where $\lambda_{|\mathcal{I}|}$ is a penalization coefficient. Suppose that $\mathcal{S}$ is the parameter space of Θ, and that

$$\hat{\Theta} = \arg\min_{\Theta \in \mathcal{S}} L(\Theta). \tag{11}$$

In practice, each pixel of a tensor image can only range from white to black and is usually normalized. Hence, it is sensible to assume that $\|\boldsymbol{\Theta}\|_\infty \leq C_0$, $\|\boldsymbol{\gamma}\|_\infty \leq C_1$ and $\|\boldsymbol{\mathcal{X}}\|_\infty \leq C_2$ for large constants $C_0 \geq 0$, $C_1 \geq 0$ and $C_2 \geq 0$. Then we define the vector parameter space $\mathcal{S}_{\boldsymbol{\Theta}} = \{\boldsymbol{\Theta} : \|\boldsymbol{\Theta}\|_\infty \leq C_0\}$ and $\mathcal{S}_{\boldsymbol{\gamma}} = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|_\infty \leq C_1\}$.

For each $X_{i,j_1\cdots j_D}^{(m)}$, let $l_\Delta(\Theta | X_{i,j_1\cdots j_D}^{(m)}) = l(\Theta, X_{i,j_1\cdots j_D}^{(m)}) - l(\Theta_0, X_{i,j_1\cdots j_D}^{(m)})$ be the loss difference, where Θ_0 corresponds to the unique true parameter. We first define

$$K(\Theta, \Theta_0) = \frac{1}{NMp_1 \cdots p_D} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j_1=1}^{p_1} \cdots \sum_{j_D=1}^{p_D} \mathbf{E}\{l_\Delta(\Theta | X_{i,j_1\cdots j_D}^{(m)})\},$$

which is the expected loss difference. Since Θ_0 is the unique true parameter, we have $K(\Theta, \Theta_0) \geq 0$ for all $\Theta \in \mathcal{S}_{\boldsymbol{\Theta}}$ and $K = 0$ if and only if $\Theta = \Theta_0$. Therefore, we define the distance between Θ and Θ_0 as $\rho(\Theta, \Theta_0) = K^{1/2}(\Theta, \Theta_0)$, and also define the variance of the loss difference as follows

$$V(\Theta, \Theta_0) = \frac{1}{NMp_1 \cdots p_D} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j_1=1}^{p_1} \cdots \sum_{j_D=1}^{p_D} \text{var}\{l_\Delta(\Theta | X_{i,j_1\cdots j_D}^{m})\}.$$

Under the L_2-loss, it is expected that $K(\Theta,\Theta_0) = \frac{1}{NMp_1\cdots p_D}\|\Theta-\Theta_0\|_F^2$, and that $V(\Theta,\Theta_0) = \frac{4\sigma^2}{NMp_1\cdots p_D}\|\Theta-\Theta_0\|_F^2$, where σ^2 is assumed to be the same variance of each element of the tensor.

Theorem 1. Suppose $\hat{\Theta}$ is the sample estimator satisfying (11), then we have

$$\mathbf{P}\left(\frac{1}{|\mathcal{I}|^{1/2}}\|\hat{\Theta}-\Theta_0\|_F \geq \tau_{|\mathcal{I}|}\right) \leq 7\exp(-c_1|\mathcal{I}|\tau_{|\mathcal{I}|}^2)$$

for $\tau_{|\mathcal{I}|} = \max(\varepsilon_{|\mathcal{I}|}, \lambda_{|\mathcal{I}|}^{1/2})$, where $c_1 > 0$ is a constant and $\varepsilon_{|\mathcal{I}|} \sim |\mathcal{I}|^{-1/2}$. The best possible rate is achieved at $\tau_{|\mathcal{I}|} = |\mathcal{I}|^{-1/2}$ when $\lambda_{|\mathcal{I}|} \sim \varepsilon_{|\mathcal{I}|}^2$.

Theorem 1 indicates that, with an appropriate rate of the penalty term shrinking to zero, the recovered tensor signal in (11) converges to the true one as the number of images goes to infinity. In other words, assuming the identifiability conditions are satisfied, Theorem 1 ensures that the extracted information from the sample images captures the true underlying information as the sample size increases. In addition, the convergence rate is established under the L_2 distance, which is a special case of the Kullback–Leibler divergence.

Remark 1. In practice, the estimator specified in (11) is a global minimizer and may not be attainable due to the nonconvex nature of the proposed optimization problem. One possible solution is to follow (1.1) of Shen (1998) and relax (11) to allow an approximate global minimizer, in the sense that the IMTL estimator approaches the global minimizer asymptotically, which is less restrictive on the requirement of the global optimum. Another solution is to develop theoretical properties directly for the IMTL estimator, which is viable because of the block-wise convexity of the proposed criterion function (Li and Zhang 2017; Wang, Zhu, and ADNI 2017). However, this may require good initial values to establish the consistency property.

Next, we investigate the theoretical property of the predictive model with the extracted sample information. In the proposed framework, we consider a generalized linear model (McCullagh and Nelder 2018) assuming that the response variable Y_i is associated to the true image signal $(\boldsymbol{\Theta}_i)$ via $\mathbf{E}(Y_i|\boldsymbol{\Theta}_i) = \mu(\eta_i)$ and $\eta_i = f(\boldsymbol{\Theta}_i)^T\boldsymbol{\beta}$, where $f : \mathbb{R}^{p_1\times\cdots\times p_D\times M} \to \mathbb{R}^{p_f}$ is the feature-extraction mapping following (6), and $\mu(\cdot)$ is a link function. Let $F(\boldsymbol{\Theta}_{|\mathcal{I}|}) = (f(\boldsymbol{\Theta}_1)^T,\ldots,f(\boldsymbol{\Theta}_N)^T)^T$ denote the extracted feature covariates based on the true imaging signals, and let $\mathcal{L}\left(y_i, f(\boldsymbol{\Theta}_i), \boldsymbol{\beta}\right) = \mathcal{L}(y_i,\eta_i\{f(\boldsymbol{\Theta}_i),\boldsymbol{\beta}\})$ denote a general loss function analogues to that in (2). Consequently, a general criterion function for supervised learning with true imaging information can be presented as

$$G_N(\boldsymbol{y}, F(\boldsymbol{\Theta}_{|\mathcal{I}|}), \boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i, f(\boldsymbol{\Theta}_i), \boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}}\cdot p^*(\boldsymbol{\beta}), \tag{12}$$

where $p^*(\boldsymbol{\beta})$ is a penalty term imposed on feature covariates effect $\boldsymbol{\beta}$, $\lambda_{\boldsymbol{\beta}}$ is the associated tuning parameter, and $\boldsymbol{y} = (y_1,\ldots,y_N)'$ is the sample response.

Note that the model in (12) essentially refers to a "true" model with the underlying true imaging information. Since only sample images with noises are observed, we actually use the recovered information $\hat{\boldsymbol{\Theta}}_{|\mathcal{I}|}$ and the corresponding sample features $F(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}) = F(\hat{\boldsymbol{\Theta}}_{|\mathcal{I}|})$ in the supervised learning model (12), leading to a sample objective function. Define $\Omega_{\boldsymbol{\beta}}$ and Ω_η as the parameter spaces for $\boldsymbol{\beta}$ and η_i $(1 \leq i \leq N)$, respectively. In the following, we introduce a regularity condition on the loss function $\mathcal{L}(\cdot)$:

(C1): For $\eta_i, \tilde{\eta}_i \in \Omega_\eta$ and all y_i's, there exists a function $K_i(\cdot)$ such that

$$|\mathcal{L}(y_i,\eta_i) - \mathcal{L}(y_i,\tilde{\eta}_i)| \leq K_i(y_i)|\eta_i - \tilde{\eta}_i|, \quad 1 \leq i \leq N,$$

and $\frac{1}{N}\sum_1^N \mathbf{E}[K_i(y_i)^2] \leq K_0$ holds for a positive constant K_0.

Lemma 2. Suppose condition (C1) holds. Under regularity condition (R1) provided in supplementary materials (A.5), and if $\frac{1}{N^{1/2}}\|F(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}) - F(\boldsymbol{\Theta}_{|\mathcal{I}|})\|_F \to 0$, then, for any given $\boldsymbol{\beta} \in \Omega_{\boldsymbol{\beta}}$, we have

$$\left|G_N(\boldsymbol{y}, F(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}), \boldsymbol{\beta}) - G_N(\boldsymbol{y}, F(\boldsymbol{\Theta}_{|\mathcal{I}|}), \boldsymbol{\beta})\right| \to_p 0.$$

Lemma 2 indicates that the sample prediction model approaches the true model asymptotically as long as the recovered imaging signal converges to the true one. Next, we investigate the asymptotic property of the sample model estimation. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$ denoting the effect of the true imaging signal, let $\hat{\boldsymbol{\beta}}(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}) = \arg\min_{\boldsymbol{\beta}} G_N(\boldsymbol{\beta}|\boldsymbol{y}, F(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}))$ be the sample estimator with observed image covariates, and let $\tilde{\boldsymbol{\beta}}(\boldsymbol{\Theta}_{|\mathcal{I}|}) = \arg\min_{\boldsymbol{\beta}} G_N(\boldsymbol{\beta}|\boldsymbol{y}, F(\boldsymbol{\Theta}_{|\mathcal{I}|}))$ be the oracle estimator with true imaging signals. Furthermore, we denote $\mathcal{R}(\hat{\mathcal{L}}_N) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(y_i, f(\boldsymbol{\mathcal{X}}_i), \hat{\boldsymbol{\beta}}(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}))$ as the empirical risk for the sample loss function, and denote $\mathcal{R}(\tilde{\mathcal{L}}_N) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(y_i, f(\boldsymbol{\Theta}_i), \tilde{\boldsymbol{\beta}}(\boldsymbol{\Theta}_{|\mathcal{I}|}))$ as the empirical risk for the oracle loss function with true imaging information. The next results provide the estimation consistency of the imaging covariates' effect as well as the empirical error.

Theorem 2. Suppose the conditions in Theorem 1 and the identifiability conditions in Proposition 1 hold. Under regularity conditions (R1)–(R4) (supplementary materials A.5) and condition (C1), if $\lambda_{\boldsymbol{\beta}} \to 0$, as $|\mathcal{I}| \to \infty$, we have

$$|\hat{\boldsymbol{\beta}}(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}) - \tilde{\boldsymbol{\beta}}(\boldsymbol{\Theta}_{|\mathcal{I}|})| \to_p 0 \quad \text{and} \quad \hat{\boldsymbol{\beta}}(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|}) \to_p \boldsymbol{\beta}_0,$$

and also,

$$\left|\mathcal{R}(\hat{\mathcal{L}}_N) - \mathcal{R}(\tilde{\mathcal{L}}_N)\right| \to_p 0.$$

In addition to model estimation consistency, Theorem 2 indicates that the empirical error of the sample predictive model approaches the oracle one, where the true imaging signals are known. With an appropriately selected predictive model in (2), the proposed approach is able to achieve a consistent and efficient prediction result.

Remark 2. The exact generalization error bound of the proposed model actually depends on a particular predictive model specified in (12) as well as the response types. Consider a new observation with true signal $(Y_j^*, \boldsymbol{\Theta}_j^*)$, in a generalized linear model,

Theorem 2 indicates that we would have a consistent mean response prediction $\hat{\mu}_j^* = \mu(f(\boldsymbol{\Theta}_j^*)^T\hat{\boldsymbol{\beta}}(\boldsymbol{\mathcal{X}}_{|\mathcal{I}|})) \to_p \mathbf{E}[Y_j^*|\boldsymbol{\Theta}_j^*]$, where $\mu(\cdot)$ is a link function. Furthermore, for a continuous and bounded response, the L_2-prediction-risk $\mathbf{E}(\hat{Y}_j^* - Y_j^*)^2$ converges to the theoretical lower bound $\text{var}_{\boldsymbol{\Theta}_j^*}(Y_j^*)$ with $\hat{Y}_j^* = \hat{\mu}_j^*$; while for a binary response Y_j^*, let $\hat{\pi}(\boldsymbol{\Theta}_j^*) = \mathbf{1}[\hat{\eta}(\boldsymbol{\Theta}_j^*) > 1/2]$ be the trained classification rule, where $\hat{\eta}(\boldsymbol{\Theta}_j^*) = \text{logit}(f(\boldsymbol{\Theta}_j^*)^T\hat{\boldsymbol{\beta}})$, then we have $\hat{\pi}(\boldsymbol{\Theta}_j^*) \to_p \pi_j^0$, the optimal Bayes' rule.

5. Simulation Studies

In this section, we investigate the performance of the proposed approach comparing with other competing methods through a simulation study of heterogeneous weak signals, which mimics the real data structure and also frequently arise in clinical diagnosis and cancer imaging analysis. Specifically, we simulate four-modality imaging data, where multiple modalities imaged from the same individual share random-location heterogeneous signals, while each modality contains its unique background bases. Additional simulations regarding various heterogeneous signal patterns with the singlemodality data and the impact of the contrast between population-shared components and individualized components are provided in the supplementary materials Section B.

In this study, we simulate the mth-modal image for the ith subject as $\boldsymbol{\mathcal{X}}_i^{(m)} = \boldsymbol{\mathcal{A}}_i^{(m)} + \boldsymbol{\mathcal{B}}_i + \boldsymbol{\mathcal{N}}_i$, $m = 1, \ldots, 4$, consisting of the true signal feature image $\boldsymbol{\mathcal{B}}_i$, the modality-specific background image $\boldsymbol{\mathcal{A}}_i^{(m)}$ and the noise background image $\boldsymbol{\mathcal{N}}_i$. For each subject, we randomly select s_i pixels of $\boldsymbol{\mathcal{B}}_i$ to be valued of 2 (signal pixels) with the other pixels of 0. We generate the response label y_i from a Bernoulli distribution with a probability of 0.4 to be 1. The number of the signal pixels s_i is generated from a Poisson distribution with means $\mu_C = 25$ and $\mu_N = 5$ for the cancer subject's image (given $y_i = 1$) and the normal subject's image (given $y_i = 0$), respectively. The noise elements of $\boldsymbol{\mathcal{N}}_i$ are generated from $N(0, 0.2^2)$.

Moreover, the first modality background $\boldsymbol{\mathcal{A}}_i^{(1)}$ is a random noise matrix with elements generated from $N(0, 0.1^2)$; the second modality has a uniform background with $\boldsymbol{\mathcal{A}}_i^{(2)} = w_i^{(2)}\mathbf{1}_D\mathbf{1}_D^T$, where $\mathbf{1}_D$ is a $D \times 1$ vector of 1's and $w_i^{(2)}$ is generated from $N(0, 0.1^2)$; and both the third and fourth modality imaging have low-rank structures with $\boldsymbol{\mathcal{A}}_i^{(m)} = \sum_{r=1}^5 w_{ir}^{(m)}\boldsymbol{a}_r^{(m),1} \circ \boldsymbol{a}_r^{(m),2}$ $(m = 3, 4)$, where $w_{ir}^{(3)}$ and $w_{ir}^{(4)}$ are generated from $N(0, 0.1^2)$, $\boldsymbol{a}_r^{(m),1}$'s and $\boldsymbol{a}_r^{(m),2}$'s are generated from $N(\mathbf{0}, \mathbf{I}_D)$ and $N(\mathbf{0}, 0.5\mathbf{I}_D)$, respectively, and $\mathbf{I}_D$ is the D-dimensional identity matrix. We set the sample size as 60 and 100, equally for the training set, the validation set, and the testing set, and the marginal imaging dimension as $D = 64$. Figure 5 illustrates sample normal and cancerous images.

To evaluate the prediction performance, we calculate the prediction accuracy rate ($\mathbf{P}[\hat{y} = y]$), the sensitivity ($\mathbf{P}[\hat{y} = 1|y = 1]$) and the specificity ($\mathbf{P}[\hat{y} = 0|y = 0]$) on the testing test. We compare the proposed IMTL model with the higher-order CP decomposition method (HOCPD) in (4), the marginal principal component analysis (MPCA) method by mimicking Caffo et al. (2010), the vectorizing L_1-penalized logistic regression model (VPL), the tensor regression model (TR, Zhou, Li, and Zhu 2013), and the CNN.

The VPL is applied on a 16,384-dimensional vector predictor by vectorizing all four modalities and then fits an L_1-penalized logistic regression model for the binary response, which is implemented by the R package "glmnet" (Friedman, Hastie, and Tibshirani 2010). The MPCA first extracts feature from each individual modality separately and then fits a logistic model with all modal features. The TR and the HOCPD are applied on the integrated multimodality image (third-order tensor predictor). The TR method is implemented by Zhou (2013)'s Matlab toolbox "TensorReg". In the following numerical studies and real data analysis, we employ an L_1-penalized logistic regression model for the IMTL, the HOCPD and the MPCA methods at the predictive stage using the extracted features. The relevant tuning parameters associated with each model are selected through minimizing the prediction error rates ($\mathbf{P}[\hat{y} \neq y]$) on the validation set, respectively.

In particular, the CNN method is implemented by Matlab toolbox "matcovnet" (Vedaldi and Lenc 2015), which inputs four modalities as four channels. Since the tuning of the CNN is crucial, we tune the CNN to minimize the classification error rates on the validation set over the number of layers, the number of filters on each convolution layer, the filter size, depth and the stride size, the pooling window size, the pooling stride size, the pooling method, the activation function, the number of epochs and the learning rates. The detailed CNN architectures and the tuning process in our numerical studies are provided in the supplementary materials Section C. In addition, we also implement the TensorFlow for the CNN and the results are consistent with the Matlab's results.

Table 1 provides the prediction results based on 100 replications. The proposed method (IMTL) outperforms other methods with the highest prediction accuracy (0.86 and 0.94) and sensitivity (0.69 and 0.88) for both sample sizes (60 and 100), respectively. Moreover, the IMTL, the HOPCD methods, and the CNN, have significant advantages over the VPL and the MPCA methods which assume the independence between the four modalities. This indicates that integrating different modalities' information at feature extraction enhances the prediction power. Furthermore, the proposed IMTL method achieves more than 16% improvement in prediction accuracy than the HOCPD, indicating that the proposed method is more effective in utilizing correlation information among different modalities with additional individualized layers. An increased training data size improves the CNN's prediction power, however, the IMTL still outperforms the CNN with a 10% higher overall accuracy and a 13.4% higher sensitivity at a sample size of 100.

Although the CNN is a powerful tool for image classification, it requires a large number of training samples to entails multiple hidden layers and involves a large number of parameters (Keshari et al. 2018; Wagner et al. 2013; Abbasi-Asl and Yu 2017a, 2017b). In practice, the sufficient sample size is not always attained, especially for cancer imaging data. In addition, the heterogeneity is another great challenge occurred in cancer imaging, while the CNN is less capable of capturing heterogeneous signals with relatively weak strength.

We conduct additional simulations to investigate the performance of the CNN compared with the proposed IMTL in

Table 1. Prediction results (SEs in subscripts) of the proposed method (IMTL) compared with five other competing methods for simulation study, based on 100 replications.

N_{tr}		VPL	TR	MPCA	HOCPD	CNN	IMTL
60	Pred accuracy	$0.593_{0.051}$	$0.573_{0.054}$	$0.598_{0.078}$	$0.731_{0.175}$	$0.714_{0.083}$	$0.858_{0.069}$
	Sensitivity	$0.000_{0.000}$	$0.415_{0.103}$	$0.443_{0.143}$	$0.526_{0.229}$	$0.379_{0.209}$	$0.687_{0.119}$
	Specificity	$1.000_{0.000}$	$0.674_{0.097}$	$0.691_{0.080}$	$0.915_{0.094}$	$0.955_{0.023}$	$0.985_{0.064}$
100	Pred accuracy	$0.608_{0.043}$	$0.603_{0.045}$	$0.634_{0.057}$	$0.808_{0.168}$	$0.884_{0.069}$	$0.941_{0.041}$
	Sensitivity	$0.001_{0.005}$	$0.453_{0.090}$	$0.490_{0.113}$	$0.682_{0.226}$	$0.774_{0.160}$	$0.878_{0.064}$
	Specificity	$0.998_{0.008}$	$0.697_{0.091}$	$0.728_{0.070}$	$0.931_{0.064}$	$0.985_{0.010}$	$0.993_{0.030}$

NOTE: Equal sample size for training set (N_{tr}), validation set (N_{vl}), and testing set (N_{ts}).

various situations. First, we compare the performance of the CNN and the IMTL, with different training sample sizes. Figure 3 shows that the IMTL consistently outperforms the CNN, especially when the training sample size is limited, while the CNN gradually "catches up" as the sample size increases. This confirms that the CNN requires a large size of training samples to perform well. The image size in this simulation is only 64×64. According to the trend in Figure 3, for real data with an image size of hundreds to thousands, the CNN could require quite a substantial number of training samples to guarantee a good performance.

Second, we investigate the performance of the CNN and the proposed IMTL with varying heterogeneous-signal strengths. To mimic real data, we utilize the concentration intensity of the TMVs to measure the strength and the heterogeneity of signals, which is controlled by the mean value of the number of the TMVs (μ_C) in cancer images, while fixing the normal images' mean ($\mu_N = 5$). Therefore, the larger the μ_C is, the stronger the signal contrast between the cancerous images and the normal images is. On the other hand, if μ_C is larger, then the heterogeneity is less since the TMVs tend to fill the entire cancer image, and thus the heterogeneity of the randomly located signals is reduced. We set the training sample size as 60. Figure 4 shows the prediction power of the CNN and the IMTL with a varying μ_C. It is clear that the IMTL outperforms the CNN consistently, while the CNN improves as the signal strength becomes stronger and the heterogeneity reduces. This simulation confirms that the proposed IMTL is more powerful than the CNN in situations where the heterogeneity is high and the signal strength is weak.

6. Real Data: Multiphoton Imaging Data for Breast Cancer

We apply the proposed method to the multimodality optical imaging data for breast cancer produced by Boppart Lab (Tu et al. 2016) at University of Illinois Urbana-Champaign. To better visualize the tissue's biological structure at cellular and molecular levels, Tu et al. (2016)'s multiphoton microscope generates multimodal images through emission of different numbers of photons. Specifically, there are two-photon auto-fluorescence (2PAF), three-photon auto-fluorescence (3PAF), second-harmonic generation (SHG), and third-harmonic generation (THG). Two-photon-fluorescence microscopy is commonly used to visualize tissue morphology and physiology at a cellular level, and three-photon-fluorescence with longer wavelength can reach deeper levels of the tissue and thus provide higher imaging resolution (Chu et al. 2005; Horton et al. 2013). This new technique is able to capture the important TMVs which are not easily identified by conventional imaging tools such as histology imaging.

Figure 6 illustrates the four modalities for a human subject's normal breast tissue and a human subject's cancerous breast tissue. In contrast to the normal tissue, multiple modalities clearly indicate a large number of TMVs which spread out in the microenvironment on the cancerous imaging, particularly

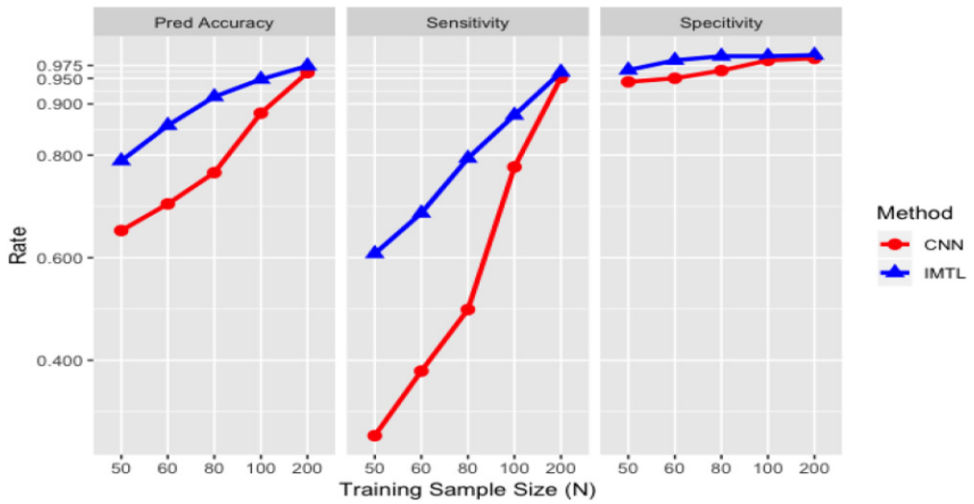

Figure 3. Prediction performance of the CNN and the IMTL under the settings of simulation study (Section 5) with different sample sizes. The results are based on 50 replications.

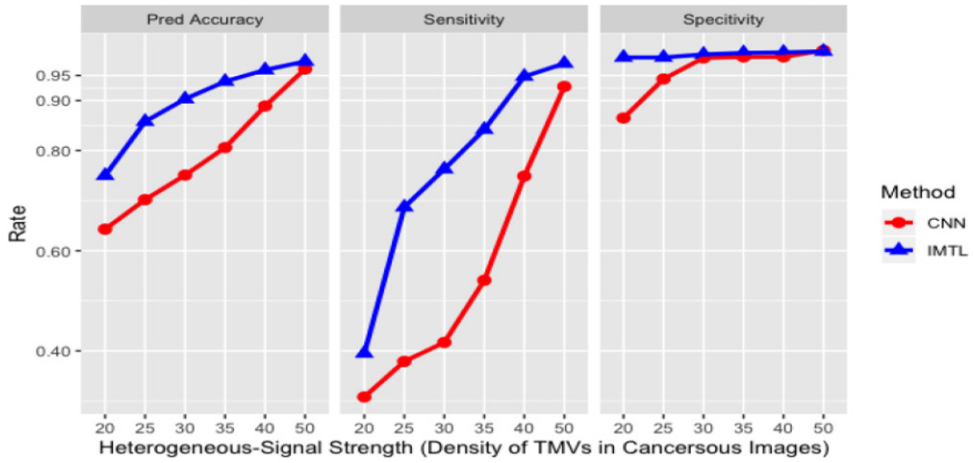

Figure 4. Prediction performance of the CNN and the IMTL under the settings of simulation study (Section 5) with different heterogeneous-signal strengths. The signal strength and the heterogeneity level are controlled by the density (mean number: μ_C) of the TMVs in a cancerous image, while the density in a normal image (μ_N) is fixed as 5. The training sample size is set as 60 and the results are based on 50 replications.

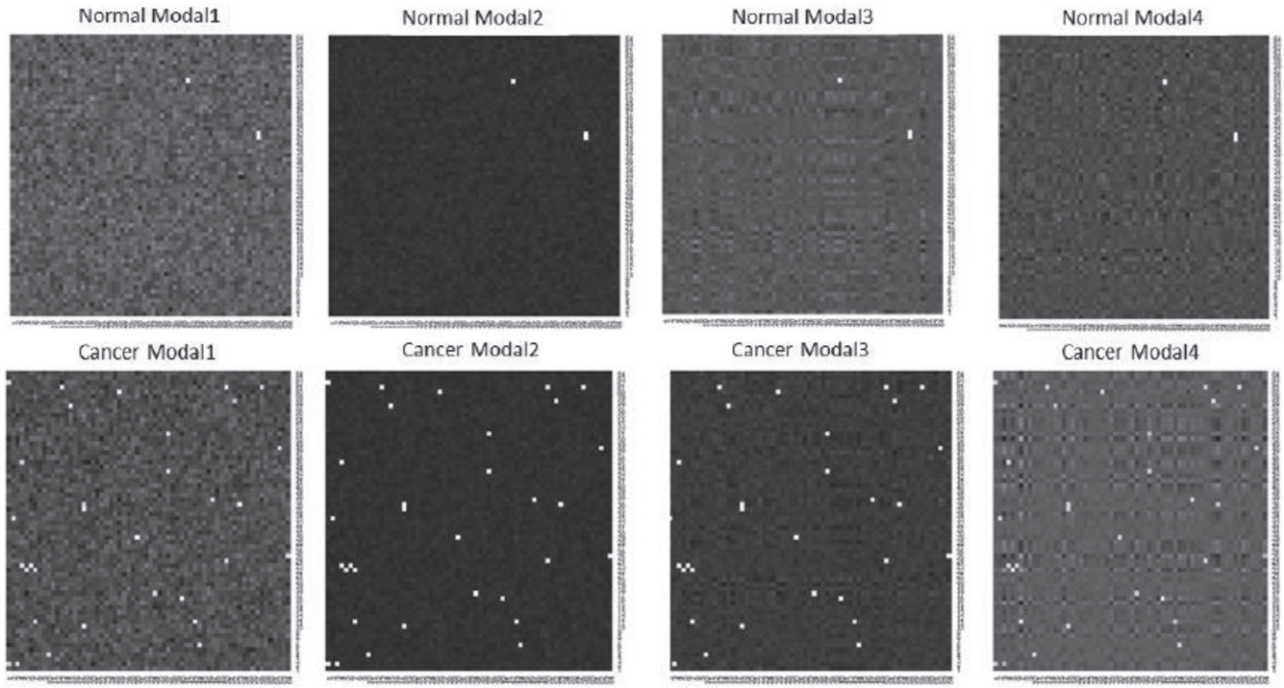

Figure 5. The representative four-modality images for a normal subject and a cancerous subject in simulation study, with each modal image size of 64 × 64.

in the 2PAF image and the 3PAF image. The normal imaging also presents microvesicles in the 3PAF image, but they are more sporadic with much less intensity.

Prior knowledge in cancer detection shows that informative TMVs are frequently observed in the microenvironment between certain biological organizations such as at the lipid boundary area and around the stroma (Tu et al. 2016). Therefore, we study segmented imaging from three normal human individuals and two cancerous individuals. Specifically, we generate sample images through segmentation, which leads to 107 normal subjects and 53 cancerous subjects, each subject having four modalities of images, with each modality of 100 × 100 pixels. To fit the predictive models and evaluate the diagnosis performance, we randomly split the total sample into a training set, a validation set, and a testing set with 60, 40, and 60 subjects, respectively. The prediction results are evaluated by prediction accuracy, sensitivity, and specificity on the testing set based on 30 replications.

We compare the proposed IMTL method to the five methods described in Section 5. Table 2 provides the averaged prediction results on the testing set, which indicates that the proposed IMTL method outperforms the other methods significantly in terms of achieving the highest overall prediction accuracy and sensitivity. Specifically, the prediction accuracy of the proposed method improves by 34%, 39%, 19%, 13%, and 7% compared to the VPL, TR, MPCA, HOCPD, and CNN approaches, respectively. The boxplot in Figure 7 suggests that the proposed method also has the smallest SE in overall prediction accuracy. Note that the outstanding predictive power of the proposed method is mainly due to its highest sensitivity

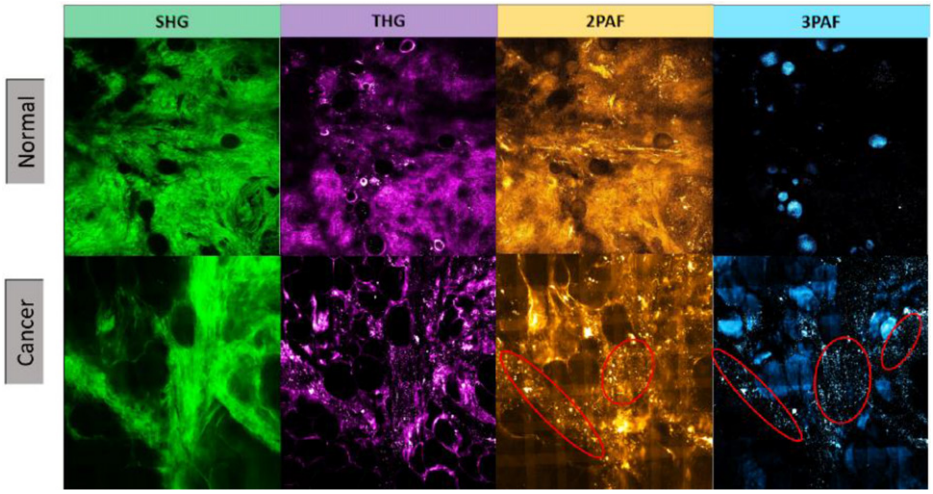

Figure 6. The four-modality microscope images for a normal human subject (2382 × 2401) and a cancerous human subject (2191 × 2193). See circles of TMVs.

Table 2. Prediction results of the proposed method (IMTL) compared with five other competing methods for human breast cancer imaging data, based on 30 random replications.

Model	VPL	TR	MPCA	HOCPD	CNN	IMTL
Pred accuracy	0.681	0.656	0.766	0.803	0.871	0.921
Sensitivity	0.086	0.628	0.535	0.656	0.729	0.903
Specificity	0.962	0.671	0.875	0.878	0.928	0.925

NOTE: Sample size for the training set, the validation set, and the testing set are 60, 40, and 60, respectively.

score, which is also of primary interest for cancer diagnosis. In practice, the proportion of the potential cancer patients is rather small compared to the general population, and correctly detecting cancer at earlier stages of cancer development is crucial and critical.

For this data, the VPL method and the tensor regression model perform inadequately compared to other methods due to the random location and the weak signal strength of the TMVs. Moreover, without incorporating individual-wise heterogeneity, the MPCA and the HOCPD are inefficient with low sensitivity. In general, the CNN provides an acceptable prediction result, however, due to limited sample size and heterogeneous imaging features, the CNN is not as effective in identifying cancerous subjects with only a 73% sensitivity rate compared to the IMTL's 90% sensitivity rate. In addition, Figure 8 provides an illustration of the prediction performance using a single modality only compared to utilizing all modalities. Figure 8 clearly shows that integrating multiple modalities improves the predictive power, especially in detecting cancerous subjects.

7. Discussion

In this article, we propose an IMTL model incorporating multimodality imaging tensor covariates to predict targeted responses. In the proposed model, we extract both individual-specific and population-wise important features simultaneously from higher-order tensor covariates through different layers, and then fit a prediction model with the extracted features. We illustrate the proposed method through numerical studies and human breast cancer imaging application on both singlemodality and multimodality data.

A major contribution of the proposed method is that we achieve heterogeneous tensor decomposition through utilizing an individualized layer in addition to population-shared modality-specific structure. Our method is motivated by a multimodality imaging study for breast cancer diagnosis, where the biomarker TMVs' are distributed with great heterogeneity and each modality has its unique background features. Most existing methods assuming homogeneous structure on signals' features are either infeasible or inefficient in our situation. In contrast, the proposed different layers are capable of capturing individual-specific spatial features through integrating different modalities' imaging information for the same individual. Both numerical studies and theoretical results demonstrate that the proposed method can achieve higher diagnostic accuracy.

In the proposed method, multilayer tensor decomposition in the first stage is not connected to the response variable directly, which may not guarantee an optimal feature extraction for supervised learning in the second stage. For future research, it is worth developing a supervised feature extraction scheme which can be more powerful in predicting outcomes. We point out three potential directions. The first one is to optimize both the classification loss and the image tensor reconstruction loss, where the latter can be treated as a part of regularization. The second possible direction is to incorporate sufficient dimension reduction techniques to search the "best" bases layers conditional on the response variable. The third direction is to link the feature extraction stage and the prediction stage in a hierarchical

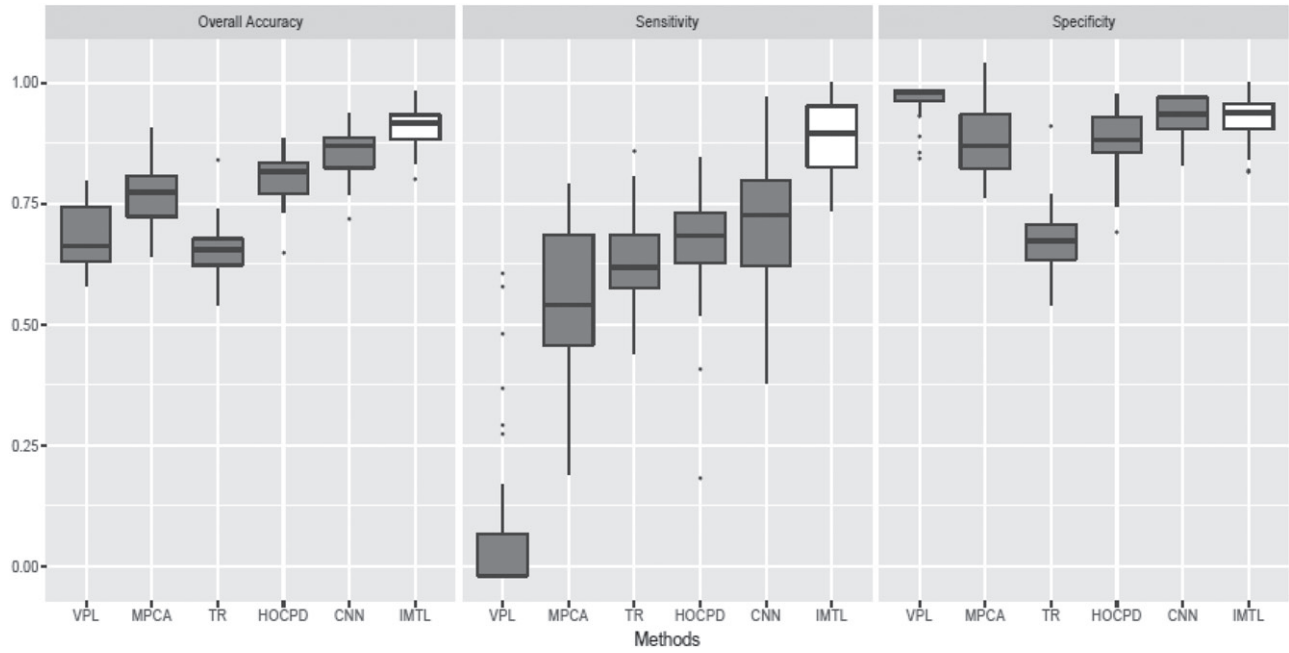

Figure 7. The boxplots of the prediction results for the VPL, the TR, the MPCA, the HOCPD, the CNN, and the IMTL models using all four modalities, based on 30 random splitting replications. Sample sizes for the training set, the validation set and the testing set are 60, 40, and 60, respectively.

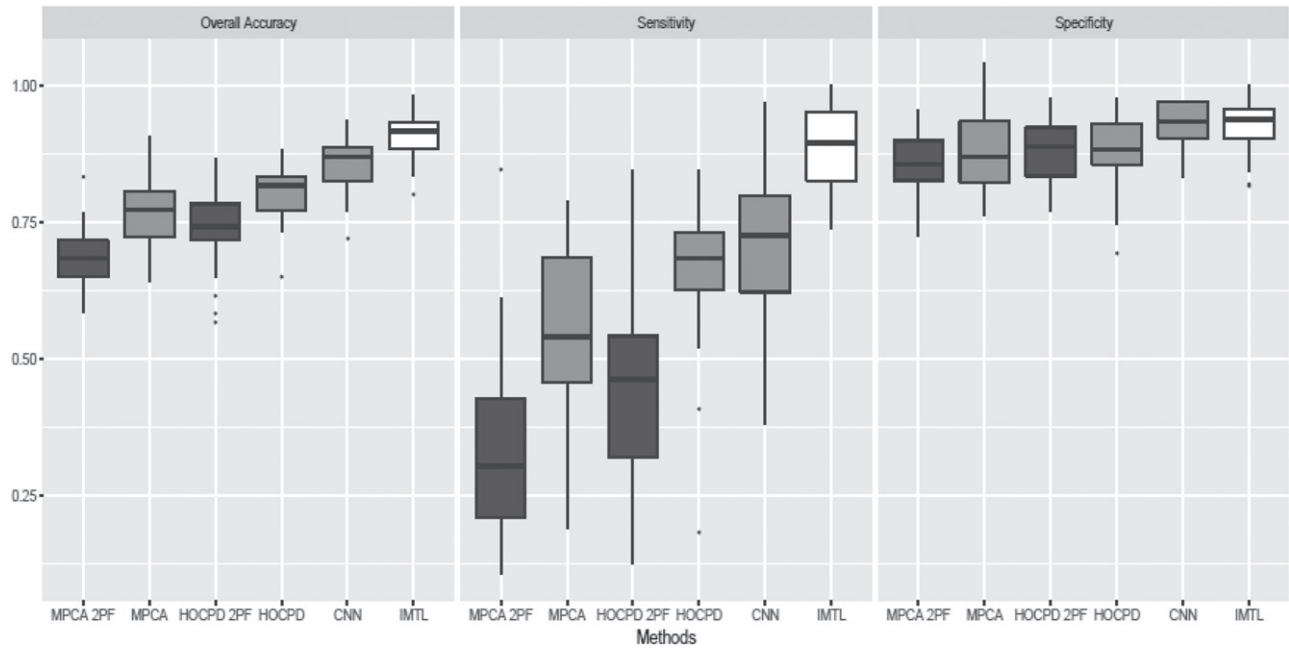

Figure 8. The boxplots of the prediction results for the MPCA model and the HOCPD model using only 2PAF modality (MPCA 2PF and HOCPD 2PF, respectively), the MPCA and the HOCPD model using all modalities, the CNN, and the IMTL models using all modalities. The results are based on 30 random splitting replications. Sample sizes for the training set, the validation set, and the testing set are 60, 40, and 60, respectively.

form mimicking the CNN framework, and update the first stage of feature extraction adaptively based on the prediction loss.

Supplementary Materials

The online supplement contains all technical proofs, additional numerical results, and computation details.

Acknowledgments

The authors are grateful to reviewers, the associate editor, and editor for their insightful comments and suggestions which have improved the manuscript significantly.

Funding

The authors would like to acknowledge support for this project from the National Science Foundation grants DMS-1415308, DMS-1613190, DMS-1821198.

References

Abbasi-Asl, R., and Yu, B. (2017a), "Interpreting Convolutional Neural Networks Through Compression," arXiv no. 1711.02329. [837,845]

——— (2017b), "Structural Compression of Convolutional Neural Networks Based on Greedy Filter Pruning," arXiv no. 1705.07356. [837,845]

Beckmann, C. F., and Smith, S. M. (2005), "Tensorial Extensions of Independent Component Analysis for Multisubject FMRI Analysis," *NeuroImage*, 25, 294–311. [837]

Bowman, F. D. (2007), "Spatiotemporal Models for Region of Interest Analyses of Functional Neuroimaging Data," *Journal of American Statistics Association*, 102, 442–453. [836]

Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. (2008), "A Bayesian Hierarchical Framework for Spatial Modeling of FMRI Data," *NeuroImage*, 39, 146–156. [836]

Caffo, B., Crainiceanu, C., Verduzco, G., Joel, S., Mostofsky, S. H., Bassett, S., and Pekar, J. (2010), "Two-Stage Decompositions for the Analysis of Functional Connectivity for FMRI With Application to Alzheimer's Disease Risk," *NeuroImage*, 51, 1140–1149. [836,837,845]

Carroll, J. D., and Chang, J. J. (1970), "Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of 'Eckart-Young' Decomposition," *Psychometrika*, 35, 283–319. [841]

Chen, B., He, S., Li, Z., and Zhang, S. (2012), "Maximum Block Improvement and Polynomial Optimization," *SIAM Journal on Optimization*, 22, 87–107. [841]

Chu, S., Tai, S., Ho, C., Lin, C., and Sun, C. (2005), "High-Resolution Simultaneous Three-Photon Fluorescence and Third-Harmonic-Generation Microscopy," *Microscopy Research and Techniques*, 66, 193–197. [846]

Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., and Schmidhuber, J. (2011), "Flexible, High Performance Convolutional Neural Networks for Image Classification," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Barcelona, Spain (Vol. 22), p. 1237. [837]

Cong, F., Phan, A. H., Zhao, Q., Huttunen-Scott, T., Kaartinen, J., Ristaniemi, T., Lyytinen, H., and Cichocki, A. (2012), "Benefits of Multidomain Feature of Mismatch Negativity Extracted by Non-negative Tensor Factorization From EEG Collected by Low-Density Array," *International Journal of Neural Systems*, 22, 1250025. [837]

Coppi, R., and Bolasco, S. (1988), "Rank Decomposition and Uniqueness for 3-Way and N-Way Arrays," in *Multiway Data Analysis*, Amsterdam: North-Holland, pp. 7–18. [838,843]

De Leeuw, J. (1994), "Block-Relaxation Algorithms in Statistics," in *Information Systems and Data Analysis*, Berlin: Springer. [841]

Derado, G., Bowman, F. D., and Kilts, C. D. (2010), "Modeling the Spatial and Temporal Dependence in FMRI Data," *Biometrics*, 66, 949–957. [836]

Fass, L. (2008), "Imaging and Cancer: A Review," *Molecular Oncology*, 2, 115–152. [836]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [845]

Friston, K. J. (2009), "Modalities, Modes, and Models in Functional Neuroimaging," *Science*, 326, 399–403. [836]

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., Johnson, S. C., and ADNI (2009), "Spatially Augmented LPboosting for AD Classification With Evaluations on the ADNI Dataset," *NeuroImage*, 48, 138–149. [836]

Hinrichs, C., Singh, V., Xu, G., Johnson, S. C., and ADNI (2011), "Predictive Markers for AD in a Multimodality Framework: An Analysis of MCI Progression in the ADNI Population," *Neuroimage*, 55, 574–589. [838]

Horton, N. G., Wang, K., Kobat, D., Clark, C. G., Wise, F. W., Schaffer, C. B., and Xu, C. (2013), "In Vivo Three-Photon Microscopy of Subcortical Structures Within an Intact Mouse Brain," *Nature Photonics*, 7, 205–209. [846]

Kang, H., Ombao, H., Linkletter, C., Long, N., and Badre, D. (2012), "Spatio-spectral Mixed Effects Model for Functional Magnetic Resonance Imaging Data," *Journal of American Statistical Association*, 107, 568–577. [836]

Karahan, E., Rojas-Lopez, P. A., Bringas-Vega, M. L., Valdes-Hernandez, P. A., and Valdes-Sosa, P. A. (2015), "Tensor Analysis and Fusion of Multimodal Brain Images," *Proceedings of the IEEE*, 103, 1531–1559. [837]

Keshari, R., Vatsa, M., Singh, R., and Noore, A. (2018), "Learning Structure and Strength of CNN Filters for Small Sample Size Training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9349–9358. [837,845]

Kolda, T. G. (2006), *Multilinear Operators for Higher-Order Decompositions* (Vol. 2), United States: Department of Energy. [838]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [838,843]

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012), "Imagenet Classification With Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105. [837]

Lange, K. (2010), *Numerical Analysis for Statisticians*, New York: Springer-Verlag. [841]

Lazar, N. (2008), *The Statistical Analysis of Functional MRI Data*, New York: Springer-Verlag. [836]

Lee, H., Kim, Y. D., Cichocki, A., and Choi, S. (2007), "Nonnegative Tensor Factorization for Continuous EEG Classification," *International Journal of Neural Systems*, 17, 305–317. [837]

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. (2015), "Spatial Bayesian Variable Selection and Grouping for High-Dimensional Scalar-on-Image Regression," *Annals of Applied Statistics*, 9, 687–713. [836]

Li, L., and Zhang, X. (2017), "Parsimonious Tensor Response Regression," *Journal of the American Statistical Association*, 112, 1131–1146. [844]

Li, X., Xu, D., Zhou, H., and Li, L. (2013), "Tucker Tensor Regression and Neuroimaging Analysis," *Statistics in Biosciences*, 10, 520–545. [837]

Lindquist, M. (2008), "The Statistical Analysis of FMRI Data," *Statistical Science*, 23, 439–464. [836]

Liu, F., Wee, C. Y., Chen, H., and Shen, D. (2014), "Inter-modality Relationship Constrained Multi-modality Multi-task Feature Selection for Alzheimer's Disease and Mild Cognitive Impairment Identification," *NeuroImage*, 84, 466–475. [838]

Luan, H., Qi, F., Xue, Z., Chen, L., and Shen, D. (2008), "Multimodality Image Registration by Maximization of Quantitative-Qualitative Measure of Mutual Information," *Pattern Recognition*, 41, 285–98. [837]

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997), "Multimodality Image Registration by Maximization of Mutual Information," *IEEE Transactions on Medical Imaging*, 16, 187–198. [837]

Martino, F. D., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008), "Combining Multivariate Voxel Selection and Support Vector Machines for Mapping and Classification of FMRI Spatial Patterns," *NeuroImage*, 43, 44–58. [836]

McCullagh, P., and Nelder, J. A. (2018), *Generalized Linear Models*, Boca Raton, FL: Routledge. [844]

Miranda, M. F., Zhu, H., and Ibrahim, J. G. (2018), "TPRM: Tensor Partition Regression Models With Applications in Imaging Biomarker Detection," *The Annals of Applied Statistics*, 12, 1422–1450. [837]

Mørup, M., Hansen, L. K., Herrmann, C. S., Parnas, J., and Arnfred, S. M. (2006), "Parallel Factor Analysis as an Exploratory Tool for Wavelet Transformed Event-Related EEG," *NeuroImage*, 29, 938–947. [837]

Nocedal, J., and Wright, S. J. (2006), *Numerical Optimization*, New York: Springer. [840]

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005), "Bayesian FMRI Time Series Analysis With Spatial Priors," *NeuroImage*, 24, 350–362. [836]

Rao, C. R., and Mitra, S. K. (1971), *Generalized Inverse of Matrices and Its Applications*, New York: Wiley. [838]

Reiss, P., and Ogden, R. (2010), "Functional Generalized Linear Models With Images as Predictors," *Biometrics*, 66, 61–69. [837]

Shen, X. (1998), "On the Method of Penalization," *Statistica Sinica*, 8, 337–357. [844]

Sidiropoulos, N. D., and Bro, R. (2000), "On the Uniqueness of Multilinear Decomposition of N-Way Arrays," *Journal of Chemometrics*, 14, 229–239. [843]

Tu, H., Liu, Y., Turchinovich, D., Marjanovic, M., Lyngsø, J. K., Lægsgaard, J., Chaney, E. J., Zhao, Y., You, S., Wilson, W., Xu, B., Dantus, M., and Boppart, S. A. (2016), "Stain-Free Histopathology by Programmable Supercontinuum Pulses," *Nature Photonics*, 10, 534–540. [846,847]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [838]

Vapnik, V., Guyon, I., and Hastie, T. (1995), "Support Vector Machines," *Machine Learning*, 20, 273–297. [838]

Vedaldi, A., and Lenc, K. (2015), "MatConvNet-Convolutional Neural Networks for MATLAB, in *Proceeding of the ACM Int. Conf. on Multimedia*. [845]

Wagner, R., Thom, M., Schweiger, R., Palm, G., and Rothermel, A. (2013), "Learning Convolutional Neural Networks From Few Samples," in *Proceedings of International Joint Conference on Neural Networks, Dallas, Texas, USA, August, 2013*. [837,845]

Wang, A., Sun, H., and Guan, Y. (2006), "The Application of Wavelet Transform to Multi-modality Medical Image Fusion," in *Networking, Sensing and Control, 2006. ICNSC'06. Proceedings of the 2006 IEEE International Conference*, pp. 270–274. [837]

Wang, L., Zhang, Z., and Dunson, D. (2017), "Common and Individual Structure of Brain Networks," arXiv no. 1707.06360. [837]

Wang, X., Zhu, H., and ADNI. (2017), "Generalized Scalar-on-Image Regression Models via Total Variation," *Journal of the American Statistical Association*, 112, 1156–1168. [844]

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., and ADNI (2012), "Multi-source Feature Learning for Joint Analysis of Incomplete Multiple Heterogeneous Neuroimaging Data," *NeuroImage*, 61, 622–632. [838]

Zhang, D., and Shen, D. (2012), "Multi-modal Multi-task Learning for Joint Prediction of Multiple Regression and Classification Variables in Alzheimer's Disease," *NeuroImage*, 59, 895–907. [838]

Zhou, H. (2013), "Matlab TensorReg Toolbox Version 0.0.2," available at *https://hua-zhou.github.io/TensorReg/*. [845]

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression With Applications in Neuroimaging Data Analysis," *Journal of American Statistics Association*, 108, 229–239. [836,837,845]