Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Correlation Tensor Decomposition and Its Application in Spatial Imaging Data

Yujia Deng, Xiwei Tang & Annie Qu

To cite this article: Yujia Deng, Xiwei Tang & Annie Qu (2021): Correlation Tensor Decomposition and Its Application in Spatial Imaging Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1938083

To link to this article: https://doi.org/10.1080/01621459.2021.1938083

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–17: Theory and Methods
https://doi.org/10.1080/01621459.2021.1938083

Taylor & Francis
Taylor & Francis Group

Correlation Tensor Decomposition and Its Application in Spatial Imaging Data

Yujia Deng[a], Xiwei Tang[b], and Annie Qu[c]

[a]Department of Statistics, University of Illinois, Urbana-Champaign, IL; [b]Department of Statistics, University of Virginia, Charlottesville, VA; [c]Department of Statistics, University of California, Irvine, CA

ABSTRACT

Multi-dimensional tensor data have gained increasing attention in the recent years, especially in biomedical imaging analyses. However, the most existing tensor models are only based on the mean information of imaging pixels. Motivated by multimodal optical imaging data in a breast cancer study, we develop a new tensor learning approach to use pixel-wise correlation information, which is represented through the higher order correlation tensor. We proposed a novel semi-symmetric correlation tensor decomposition method which effectively captures the informative spatial patterns of pixel-wise correlations to facilitate cancer diagnosis. We establish the theoretical properties for recovering structure and for classification consistency. In addition, we develop an efficient algorithm to achieve computational scalability. Our simulation studies and an application on breast cancer imaging data all indicate that the proposed method outperforms other competing methods in terms of pattern recognition and prediction accuracy.

ARTICLE HISTORY
Received June 2020
Accepted May 2021

KEYWORDS
Dimension reduction; Image processing; Multidimensional data; Spatial correlation; Tensor decomposition

1. Introduction

Recent advances in technology have catalyzed the rapid growth of large volumes of biomedical data with heterogeneity structures. Among them, medical imaging data are the most distinctive and progressive due to their sheer volume and the importance of the fields, as medical imaging data can carry significant amounts of information on disease status and treatment outcomes.

This article is motivated by the multiphoton optical imaging data arising from a breast cancer study (Tu et al. 2016), where multiple modalities of images are taken at each target region with respect to different photon wavelengths. This advanced technology is capable of capturing tumor-associated microvesicles (TMVs) which have been shown a biomarker in detecting early-stage breast cancer before a tumor forms (D'Souza-Schorey and Clancy 2012). In particular, the number of TMVs is highly correlated with tumor aggressiveness and metastatic phenotype (Taylor and Gercel-Taylor 2008). A large quantity of spatially concentrated TMVs is a strong indicator of invasive tumors in a later stage.

Moreover, different from other imaging data such as brain imaging (Bowman, Guo, and Derado 2007; Lindquist 2008; Tian 2010) where the regions of interest are fixed, the location of tumor-associated TMVs on breast cancer images are random. Due to the randomness of signal regions and large-volume of pixels, traditional regression methods assuming fixed-location signals are likely to fail in breast cancer imaging. Instead, Tang, Bi, and Qu (2019) considered heterogeneous structures among subjects and proposed to extract features based on individualized tensor decomposition. However, one drawback of their method is that it is hard to interpret the relationship between the extracted features and the label of individuals, and therefore it is rather difficult to recover the signal regions that contain TMVs. In addition, Tang, Bi, and Qu (2019) did not utilize the spatial correlation information among pixels to provide more effective detection on TMV and improve classification accuracy.

In this article, we propose to incorporate pixel correlation in multimodality image analysis to identify TMVs more efficiently. The existing approaches in multimodality image analysis (Hinrichs et al. 2011; Yuan et al. 2012; Zhang and Shen 2012; Liu and Calhoun 2014; Tang, Bi, and Qu 2019) are mainly based on the marginal mean of imaging pixels. However, these methods are not applicable and likely to fail when the signals are too weak compared to noisy backgrounds, which does not provide informative features for imaging diagnosis. In addition, the pixel values might present heterogeneously across different modalities. In particular, signals can be more visually expressed in certain modalities while less apparent in others, which also makes it more difficult to detect based on pixel values only. Nevertheless, the signals still share certain imaging features across different modalities; particular spatial correlations are still more or less preserved even under different modalities. This motivates us to consider a new tensor correlation strategy to incorporate correlation structures of pixels for multimodality images.

However, the estimation of the correlation structure is non-trivial given that image size is much larger than modality numbers. Existing approaches for handling high-dimensional correlation matrices include the matrix normal distribution (Dutilleul 1999) which assumed that the covariance of the vectorized data can be decomposed into a Kronecker product of two matrices. Nonetheless, the decomposed covariance matrices

CONTACT Annie Qu aqu2@uci.edu Department of Statistics, University of California, Irvine, CA 92697.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

can only capture the marginal correlation between rows and columns of the matrix, and cannot represent the neighboring spatial information. The principal orthogonal complement thresholding method (POET) Fan, Liao, and Mincheva (2013) assumed a factor model with a sparse error covariance matrix, but it is not applicable for imaging data where the factors are no longer single dimensions and the error covariance is not sparse.

Detection of TMVs can be formulated as a clustering problem through the correlation structure of pixels. The existing clustering approaches require index permutation marginally (DeRisi, Iyer, and Brown 1997; Eisen et al. 1998; Bar-Joseph, Gifford, and Jaakkola 2001), where the neighborhood information is ignored. However, spatial information is crucial in detecting TMVs, otherwise irrelevant pixels can be classified as TMVs incorrectly, which may result in high false positive rate. Moreover, correlation estimation after vectorization is inefficient as it only utilizes information of single-pixel pairs without incorporating high-order spatial information jointly in multimodality image analysis.

To tackle the above challenges, we introduce a new concept of correlation tensor which captures the spatial correlations for multimodality imaging data in a high-order array. In addition, we propose a novel semi-symmetric tensor decomposition method to recover the correlation structures among all pixels efficiently and extract latent features for disease diagnosis. Specifically, we impose a low-rank structure in tensor decomposition to reduce the number of parameters, which enables us to detect signal patterns on images associated with potential disease outcomes.

A low-rank structure has been adopted in many tensor models to discover the underlying features. For example, Bi, Qu, and Shen (2018) extracted the subgroup information through tensor decomposition to improve the recommendation system; Tang, Bi, and Qu (2019) developed an individualized multilayer tensor learning method to classify multimodal images; Zhang and Xia (2018) investigated the consistency of tensor singular value decomposition (SVD) under different signal noise ratios; Xia and Zhou (2019) proposed a low-rank tensor denoising estimator with sharp entry-wise deviation bounds; Sun and Li (2019) utilized the low-rank tensor factorization to achieve dynamic tensor clustering. In addition, Allen (2012) obtained both low-rank and sparse tensor decomposition by imposing L_1 penalty on components, and Zhang and Han (2019) achieved sparse tensor SVD through iterative thresholding. However, all existing tensor methods assume that the elements of the target tensor are independent, while in our case, the entries of the decomposed tensors are sample correlation estimates and are correlated in nature. The dependence nature imposes a great challenge in deriving theoretical properties compared to the independent case.

To the best of our knowledge, the proposed method is the first to integrate correlation structure and spatial information under the tensor framework. Through incorporating spatial correlation information, the proposed method is able to provide higher classification accuracy when signal strength is weak compared with background noise. Since spatial information is preserved, we are able to fully use the multi-dimensional structure of image data, and distinguish signals from noise according to their spatial distribution in the image, which is infeasible using traditional vectorization methods. In addition, we are also able to identify regions that contain signals associated with tensor correlation features directly, which provides more meaningful scientific interpretation compared to Tang, Bi, and Qu (2019). Moreover, the semi-symmetric tensor decomposition improves estimation efficiency by reducing the dimension of parameters, which enables us to extract important features even with a limited number of modalities. More importantly, we establish a general theoretical framework using the low-rank tensor decomposition model for high-dimensional correlated data which takes underlying correlations into account. The new developed theory generalizes the theoretical properties under the independence framework Wang and Li (2020) to accommodate different correlation structures and regularization terms.

Our numerical results and theoretical properties all indicate that the proposed method achieves higher classification accuracy with an increase of the number of modalities and image size. In the existing regression methods Li et al. (2018a), the increase of modalities and image size adds more difficulties to prediction since it introduces more coefficients to estimate. In contrast, the proposed method utilizes additional information on spatial correlation and multimodality for better feature extraction through correlation tensor decomposition, which leads to better prediction performance.

The rest of the article is organized as follows. Section 2 introduces notations and some background in tensor analysis. Section 3 proposes the correlation tensor structure, the semi-symmetric decomposition method and the corresponding classification method based on the extracted features. Section 4 provides the identifiability and asymptotic result of the proposed method. Section 5 demonstrates the numerical studies using simulated data. Section 6 applies the proposed method to multimodal breast cancer imaging data. The last section provides concluding remarks and discussion.

2. Notation and Background

We start with some notations to represent multimodality image data. Let $X^{(i,m)}$ be the observed imaging data for the ith subject on the mth modality, where $1 \le i \le n$ and $1 \le m \le M$. We assume that each $X^{(i,m)}$ is an independent and identically distributed sample of the random matrix $X^{(i)}$. For simplicity, we assume that the image size is $L \times L$ since the generalization to the rectangular case is straightforward. We denote $X^{(i)}_{pq}$ as the (p,q)th pixel of $X^{(i)}$ with mean $\mathbb{E}(X^{(i)}_{pq}) = \mu^{(i)}_{pq}$, marginal variance $\text{var}(X^{(i)}_{pq}) = (\sigma^{(i)}_{pq})^2$, $(p,q = 1,\ldots,L)$; and the correlation with another pixel $X^{(i)}_{st}$ as $\text{corr}(X^{(i)}_{pq}, X^{(i)}_{st}) = \rho^{(i)}_{pqst}$. In this study, we focus on the case when $\mu^{(i)}_{pq}$ is noisy or noninformative and we utilize the correlation pattern $\rho^{(i)}_{pqst}$ to classify images.

One conventional representation of the correlation structure is based on the vectorized data Dutilleul (1999), Manceur and Dutilleul (2013), and Hoff (2011), that is, $\Sigma_0 = \text{corr}\{\text{vec}(X)\}$ where $\Sigma_{0,q+(p-1)L,s+(t-1)L} = \text{corr}\left(X_{pq}, X_{st}\right)$. However, one disadvantage of vectorizing is that it cannot preserve the important spatial information which is crucial in detecting the target-

ing signal patterns associated with spatially correlated pixels. Another issue of vectorization is the estimation efficiency of the sample correlation matrix, since the number of parameters involved through vectorization $\mathcal{O}(L^4)$ is much larger than the number of modalities M.

In this article, we propose to use tensor structure to represent the correlation information. We provide some background on tensor and its operations as follows. A Dth-order tensor is a D-dimensional array $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots p_D}$, where the order of a tensor is defined as the number of dimensions, also known as ways or modes (Kolda 2006). For example, a vector $\boldsymbol{a}$ is the first-order tensor and a matrix A is a second-order tensor. We use $\mathcal{X}_{i_1 \ldots i_D}$ to denote the $(i_1, i_2, \ldots, i_D)$th element of a tensor $\mathcal{X}$, $i_d = 1, \ldots, p_d$.

In addition, a *fiber* is defined by fixing every index of the tensor modes except one Kolda and Bader (2009). For example, for a matrix A, the ith row fiber $A_{i:}$ and the jth column fiber $A_{:j}$ correspond to the ith row vector and the jth column vector of A, respectively. A *slice* is a two-dimensional representation of a tensor, defined by fixing all except two indices. Furthermore, we define the *blocks* of a Dth-order tensor with size K as the index set $\{i_1, i_1+1, \ldots, i_1+K-1\} \times \cdots \times \{i_D, i_D+1 \ldots, i_D+K-1\}$, where $\times$ denotes the Cartesian product.

In contrast to matrices, the definition of symmetry and diagonal can be ambiguous in high-order tensors. Conventionally, a tensor is called *supersymmetric* if its elements remain constant under any permutation of the indices Kolda and Bader (2009), i.e. $\mathcal{X}_{i_1 \ldots i_D} = \mathcal{X}_{i_{\sigma_1} \ldots i_{\sigma_D}}$ for every permutation σ of the symbol $\{1, 2, .., D\}$. Also, a tensor can be partially *symmetric* in two or more modes, if by fixing the rest of the indices are its slices symmetric. For example, a fourth-order tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3 \times p_4}$ is partially symmetric in mode one and two if $p_1 = p_2$ and

$$\mathcal{X}_{::st} = \mathcal{X}_{::st}^T, \quad \text{for all } s = 1, \ldots, p_3 \text{ and } t = 1, \ldots, p_4.$$

On the other hand, we call the entries $\mathcal{X}_{i_1, \ldots i_D}$ *primal diagonal* terms if $i_1 = \ldots = i_D$. For a fourth order tensor with $D = 4$, we define $\mathcal{X}_{pqst}$ as *diagonal* terms if $(p, q) = (s, t)$, and the rest as off-diagonal terms.

In the following, we introduce some tensor operations. We use $\text{vec}(\mathcal{X})$ to denote the vectorization of a tensor $\mathcal{X}$, where the $\left\{ i_1 + \sum_{d=2}^{D} \left\{ (i_d - 1) \prod_{j=1}^{d-1} p_j \right\} \right\}$-th element of $\text{vec}(\mathcal{X})$ corresponds to $x_{i_1, \ldots, i_D}$. Moreover, an outer product "$\circ$" on multiple vectors $\boldsymbol{b}_1 \in \mathbb{R}^{p_1}, \ldots, \boldsymbol{b}_D \in \mathbb{R}^{p_D}$ creates a rank-1 tensor $\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D$ where $x_{i_1 \ldots i_D} = \boldsymbol{b}_{1, i_1} \boldsymbol{b}_{2, i_2} \ldots \boldsymbol{b}_{D, i_D}$.

Hence, a Dth-order tensor $\mathcal{X}$ is defined as a rank R if it can be represented as

$$\mathcal{X} = \sum_{r=1}^{R} \boldsymbol{b}_1^{(r)} \circ \boldsymbol{b}_2^{(r)} \circ \cdots \circ \boldsymbol{b}_D^{(r)},$$

where $\boldsymbol{b}_d^{(r)}$'s $(r = 1, \ldots, R)$ are the p_d-dimensional factor vectors $(d = 1, \ldots, D)$. The above decomposition is called the CANDECOMP/PARAFAC (CP) decomposition Hitchcock (1927) which is adopted in the following sections since the rank of CP decomposition is better defined compared to the Tucker decomposition Tucker (1966).

3. Methodology

3.1. Correlation Tensor Decomposition

In this section, we introduce a new concept of correlation tensor, and the corresponding decomposition method to analyze spatial correlated multimodal imaging data.

Definition 1. For a random matrix $X \in \mathbb{R}^{L \times L}$, a *correlation tensor* is the fourth-order tensor $\mathcal{T} = \{\tau_{pqst}\}_{1 \le p,q,s,t \le L} \in \mathbb{R}^{L \times L \times L \times L}$, where $\tau_{pqst} = \text{cov}(X_{pq}, X_{st})$; and a *correlation tensor* is the fourth-order tensor $\mathcal{C} = \{c_{pqst}\}_{1 \le p,q,s,t \le L} \in \mathbb{R}^{L \times L \times L \times L}$, where $c_{pqst} = \text{corr}(X_{pq}, X_{st})$.

In practice, the pixels of an image are usually normalized so that the variances are the same. Thus, the correlation tensor and the correlation tensor are identical up to a constant. For ease of notation, we only use $\mathcal{C}$ for the rest of the article.

An illustration of the fourth-order tensor is provided on the left side of Figure 1, where the entire tensor is a set of $L \times L$ slices with an $L \times L$ matrix. The advantage of using the correlation tensor is to facilitate the correlation structure analysis while preserving the spatial information. Specifically, for an image with spatially correlated pixels located in a size K block region $\{i, \ldots, i+K-1\} \times \{j, \ldots, j+K-1\}$, where the nonzero correlation coefficients correspond to a block region $\{i, \ldots, i+K-1\} \times \{j, \ldots, j+K-1\} \times \{i, \ldots, i+K-1\} \times \{j, \ldots, j+K-1\}$ in the correlation tensor. The ultimate goal of our study is to detect the block-wise correlated pixels by identifying the block structure of the correlation tensor.

Under the above block structure assumption, we assume the correlation tensor can be approximated by a low-rank tensor decomposition

$$\mathcal{C} = \sum_{r=1}^{R} \boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)} \circ \boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)}, \quad \text{for all off-diagonal terms,} \quad (1)$$

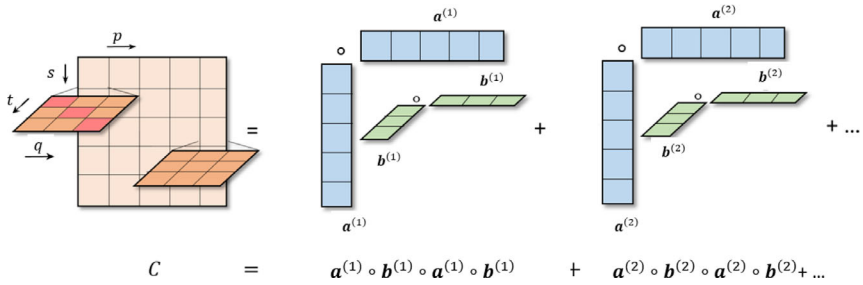

Figure 1. Illustration of the correlation tensor decomposition.

where $\boldsymbol{a}^{(r)} \in \mathbb{R}^L, \boldsymbol{b}^{(r)} \in \mathbb{R}^L$, and R is the rank. Equivalently, for each element, we have

$$c_{pqst} = \sum_{r=1}^{R} a_p^{(r)} b_q^{(r)} a_s^{(r)} b_t^{(r)}, \quad \text{for any} (p,q) \neq (s,t).$$

We denote the column stacks of $\boldsymbol{a}^{(r)}$'s and $\boldsymbol{b}^{(r)}$'s as $A = [\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \cdots, \boldsymbol{a}^{(r)}] \in \mathbb{R}^{L\times R}$, $B = [\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \cdots, \boldsymbol{b}^{(r)}] \in \mathbb{R}^{L\times R}$, respectively.

Compared with the traditional CP decomposition, the proposed decomposition (1) imposes symmetry on the decomposed bases. This is due to the symmetry of the correlation coefficients in the sense that $c_{pqst} = c_{stpq}$, for any p, q, s, t. Notice that this type of symmetry structure in the tensor is different from the *symmetric* and *partially symmetric* tensors defined earlier, and are not considered in existing tensor decomposition methods, such as Allen (2012), Shah, Rao, and Tang (2015), and Sun et al. (2017). We name (1) *semi-symmetric* tensor decomposition.

Compared to vectorization methods, the main advantage of the semi-symmetric tensor decomposition (1) is to capture the spatial correlation efficiently by reducing the dimension of parameters. This is because the vectorized approaches estimate the correlation matrix through every pairwise sample correlation $\text{corr}(X_{pq}, X_{st})$ independently, without borrowing information from other pixels. In contrast, the proposed method estimates the corresponding pairwise correlation based on the basis vectors $\boldsymbol{a}^{(r)}$'s and $\boldsymbol{b}^{(r)}$'s, which allows integrating information from all neighboring image pixels. Indeed, the number of parameters in the proposed model is reduced from $(L^2-1)(L^2)/2$ to $2LR$ compared with the traditional vectorized approaches. Therefore, we can achieve more accurate estimation on correlation coefficients, and gain more power in identifying spatial correlation.

Since the true value of $\mathcal{C}$ is unknown, we plug in the sample correlation tensor to estimate the basis vectors $\boldsymbol{a}^{(r)}$ and $\boldsymbol{b}^{(r)}$ as follows

$$\begin{aligned} &\min_{\mathcal{C}} \|\bar{\mathcal{C}} - \mathcal{C}\|_F^2 + \lambda J(\mathcal{C}), \\ &\text{subject to } \mathcal{C} = \sum_{r=1}^{R} \boldsymbol{a}_i^{(r)} \circ \boldsymbol{b}_i^{(r)} \circ \boldsymbol{a}_i^{(r)} \circ \boldsymbol{b}_i^{(r)}, \\ &\text{for all off-diagonal components,} \end{aligned} \tag{2}$$

where $J(\mathcal{C})$ is a penalty term, λ is the penalization parameter and $\bar{\mathcal{C}} = \{\bar{c}_{pqst}\}_{1\le p,q,s,t\le L}$ is the sample correlation tensor

$$\bar{c}_{pqst} = \frac{1}{M} \sum_{m=1}^{M} X_{pq}^{(m)} X_{st}^{(m)}. \tag{3}$$

In the target function (2), we use $\|\cdot\|_F^2$ to denote the summation of squares over all elements in the tensor.

We remark that the proposed framework targets a general form of correlation tensor decomposition assuming a low-rank structure. However, in some specific applications, we also allow an additional structure pursuit on the decomposed factors by including a corresponding regularization term $J(\mathcal{C})$. In general, the penalty is imposed on the decomposed factors A and B, and thus $J(\mathcal{C}) = J_1(A) + J_2(B)$. For example, a Lasso penalty for sparsity Allen (2012), a fusion penalty imposed for local smoothing (Tibshirani et al. 2005; Sun and Li 2019; Wu et al. 2019) or a graphical Lasso to encourage certain geometric structures (Madrid-Padilla and Scott 2017; Greenewald, Zhou, and Hero 2019). For ease of notation, we use $J(\mathcal{C})$ in the following to represent a general penalty form.

In the above discussion, we view multimodality images as iid observations of a single subject. However, we may also consider the case where modalities are not independent with each other. Specifically, we replace $\bar{\mathcal{C}}$ in the target function (2) with a generalized estimate of correlation that is the solution to an estimation equation incorporating cross-modality correlation. A detailed discussion and the corresponding numerical results can be found in the supplementary materials.

In addition, the decomposition enables us to detect the block-wise correlated pixels through the outer-product $\boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)}$. For example, the simplest case consists of a single block of size $s_1 \times s_2$ on the top-left corner of a random matrix $X \in \mathbb{R}^{L\times L}$, $s_1 \le L$, $s_2 \le L$. Suppose the pixels within a block are correlated with an equal correlation coefficient ρ, and uncorrelated with the rest of the pixels otherwise, that is,

$$\text{corr}(X_{pq}, X_{st}) = \begin{cases} \rho, & \text{if } p,s \le L \text{ and } q,t \le L, \\ 0, & \text{otherwise,} \end{cases}$$

then the correlation tensor can be written as $\mathcal{C} = \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{a} \circ \boldsymbol{b}$, $\boldsymbol{a} \in \mathbb{R}^L, \boldsymbol{b} \in \mathbb{R}^L$ for all off-diagonal terms, where $\boldsymbol{a} = \rho_a(\underbrace{1,\ldots,1}_{s_1},0,\ldots,0)$, $\boldsymbol{b} = \rho_b(\underbrace{1,\ldots,1}_{s_2},0,\ldots,0)$ and $\rho_a^2\rho_b^2 = \rho$. Therefore, the nonzero terms of $\boldsymbol{a} \circ \boldsymbol{b}$ correspond to the block-wise correlated region of X.

The above example indicates that the interaction term $\boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)}$ is important in identifying signals. For this purpose, we define the latent feature $F \in \mathbb{R}^{L\times L}$ as

$$F_{L\times L} = \text{abs}(A)_{L\times R}\, \text{abs}(B)_{L\times R}^T, \tag{4}$$

where $\text{abs}(\cdot)$ is applied element-wise to avoid the indeterminacy caused by sign flipping. The latent feature F has the same dimension as the original image X and its nonzero entries imply the signal region, which plays an important role in signal detection in Section 3.2.

An example. To better illustrate the idea of correlation tensor decomposition, we use the following toy example for demonstration. We generate 10 independent and identically distributed random matrices $X^{(m)} \in \mathbb{R}^{100\times 100}, m = 1,\ldots,10$, where $\text{vec}(X^{(m)})$ follows a normal distribution with mean $\mathbf{0}$ and marginal variance =1. The correlation tensor $\mathcal{C} = \sum_{r=1}^{2} \boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)} \circ \boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)}$ for all off-diagonal terms, where $\boldsymbol{a}^{(1)} = \left(\sqrt{0.9}\mathbf{1}_{50}^T, \mathbf{0}_{50}^T\right)^T$, $\boldsymbol{a}^{(2)} = \left(\mathbf{0}_{50}^T, \sqrt{0.9}\mathbf{1}_{50}^T\right)^T$, $\boldsymbol{b}^{(1)} = \left(\mathbf{0}_{75}^T, \sqrt{0.8}\mathbf{1}_{25}^T\right)^T$, $\boldsymbol{b}^{(2)} = \left(\sqrt{0.8}\mathbf{1}_{25}^T, \mathbf{0}_{75}^T\right)^T$,

Figure 2 shows four observations of the simulated random matrix $X^{(m)}$s. The generated matrices display two regions of highly correlated pixels: one is located in the lower-left and the other one is in the top-right. With the correlation tensor decomposition, we recover the underlying block-wise correlated structure successfully as shown in Figure 3.

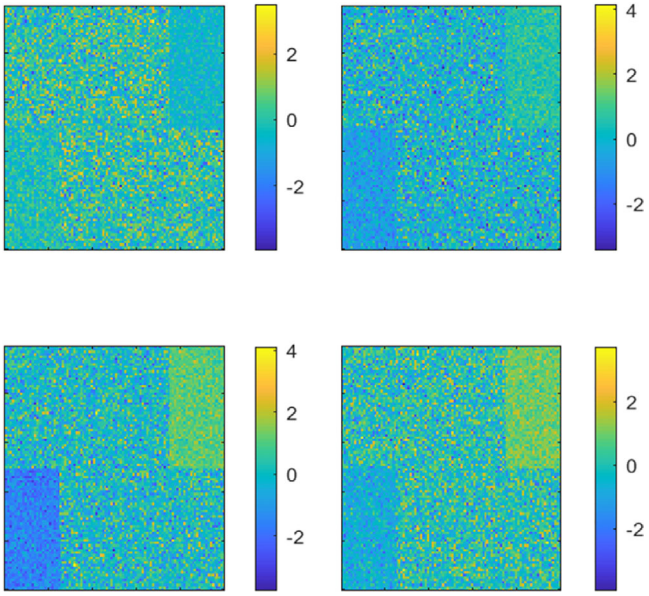

Figure 2. Simulated data from the example of Section 3.1, showing four modalities.

3.2. Latent Feature Extraction for Image Classification

In this section, we link the correlation tensor structure to image classification based on the recovered spatial-correlated feature map.

A unique advantage of the proposed decomposition method is that the spatial information in X is preserved in the latent feature F defined in Equation (4). To use this property in image classification, we extract the spatial information by applying a local operator and a global pooling function. Specifically, we define $\boldsymbol{\phi}_\beta : \mathbb{R}^{L\times L} \to \mathbb{R}^{(L-\beta+1)\times(L-\beta+1)}$ elementwise by

$$\{\boldsymbol{\phi}_\beta(F)\}_{p,q} = \frac{1}{\beta^2}\sum_{u=1}^{\beta}\sum_{v=1}^{\beta} f^2_{p+u-1,q+v-1}, \text{ for any } 1 \le p \le L-\beta+1,\ 1 \le q \le L-\beta+1,$$

where $0 < \beta \le L$ is a positive integer. And we define

$$\pi_\beta(F) = \max \boldsymbol{\phi}_\beta(F).$$

Intuitively, $\boldsymbol{\phi}_\beta$ computes the average of the square sum within a $\beta \times \beta$ block of F and π_β computes the maximum value over all blocks. By computing the maximum value, we are able to identify the signal block regardless of its relative location in the image as long as the non-zero elements within the signal block are distinguishable. The entire procedure is analogous to the convolution operator and the max pooling operator in the CNN.

Motivated by the breast cancer diagnosis application, we classify subjects with multimodality images based on the magnitude of correlations and spatial concentrations of the highly correlated pixels. We assume that the label of the ith subject Y_i is determined by

$$Y_i = \text{sign}\left\{\pi_{\beta_0}(F_i) - \delta_0\right\}, \tag{5}$$

where $0 < \delta_0 < 1$ and $0 < \beta_0 < L$ are two thresholding parameters for the correlation strength and signal area, respectively, and $\text{sign}(x) = 1$ if $x > 0$ and -1 if $x < 0$. That is, a subject is classified as cancerous if it contains at least one $\beta_0 \times \beta_0$ block such that the average value of the correlation-based latent features within this block is greater than a positive threshold δ_0.

We train the classifier through the observed multimodal images from n subjects. Specifically, we denote $\mathcal{L}(y_i, F_i, \beta, \delta) = \mathbb{1}\left[y_i \ne \text{sign}\{\pi_\beta(F_i) - \delta\}\right]$ as the 0–1 loss function, and solve for β, δ by minimizing the empirical risk function

$$(\hat{\beta}, \hat{\delta}) = \arg\min_{(\beta,\delta)} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \hat{F}_i, \beta, \delta),$$

where $\hat{F}_i$ is the estimated latent feature of the ith subject. In particular, we train β and δ by grid search on $[0, L]$ and $[0, 1]$, respectively. The flowchart of the above classification procedure is illustrated in Figure 4.

In summary, the proposed method builds a classification based on the images' spatial correlation information instead of the marginal intensity of pixels as in conventional classification methods (Caffo et al. 2010; Zhou, Li, and Zhu 2013), which makes it robust for the case when the signal strength of a single image is weak and noisy. Moreover, the proposed method does not require preregistration of images since it can accommodate heterogeneous imaging data with random locations. When the observed signal region does not have a block shape, the selected features are still able to approximate the target region since the proposed decomposition captures the principal correlation structure. This is also supported numerically in Section 5.2.

3.3. Algorithm and Implementation

For a better illustration, in this section, we introduce an efficient algorithm to solve the proposed semi-symmetric tensor decomposition (2) with an L_1-penalty $J(\mathcal{C}) = \lambda\sum_{r=1}^{R}\left|\boldsymbol{a}^{(r)}\right|_1 + \left|\boldsymbol{b}^{(r)}\right|_1$, which is commonly used for sparsity pursuit in many applications. However, the proposed framework allows a general class of penalties, and our algorithm can be easily extended to accommodate other L_p regularizations.

The semi-symmetric structure brings an additional challenge in the implementation. The traditional CP decomposition Hitchcock (1927) adopted an alternating updating strategy in computation, where the estimation of factor parameters of each mode is equivalent to a least-square problem, and has an explicit solution. In contrast, in our case, the loss function is a fourth-order polynomial due to the symmetry of $\boldsymbol{a}^{(r)}$ and $\boldsymbol{b}^{(r)}$, and there is no direct solution. Although the gradient based method can be used in each updating step, it is inefficient when the number of iterations is large and the dimension is high.

Instead of solving for $\boldsymbol{a}^{(r)}$ and $\boldsymbol{b}^{(r)}$ directly, we consider solving $\boldsymbol{a}^{(r)} \circ \boldsymbol{a}^{(r)}$ and $\boldsymbol{b}^{(r)} \circ \boldsymbol{b}^{(r)}$ first, and then obtain $\boldsymbol{a}^{(r)}$ and $\boldsymbol{b}^{(r)}$ by performing constrained singular value decomposition. Specifically, we let $\mathcal{A} \in \mathbb{R}^{L\times L\times R}$ be a third-order array where $\mathcal{A}_{psr} = a^{(r)}_p a^{(r)}_s$, and the slice $\mathcal{A}_{::r}$ is equal to $\boldsymbol{a}^{(r)} \circ \boldsymbol{a}^{(r)}$. We denote

$$D^{(ps)} = \text{Diag}\left\{(A_{p1}A_{s1}, \ldots, A_{pR}A_{sR})\right\} = \text{Diag}(\mathcal{A}_{ps:}),$$

where $\text{Diag}(\boldsymbol{x})$ denotes a matrix with the diagonal terms equal to $\boldsymbol{x}$. We ignore the penalty term first and reformulate the loss function as

$$L(A, B) = \sum_{1\le p\le s\le p_1} \|\bar{\mathcal{C}}_{p:s:} - BD^{(ps)}B^T\|_F^2,$$

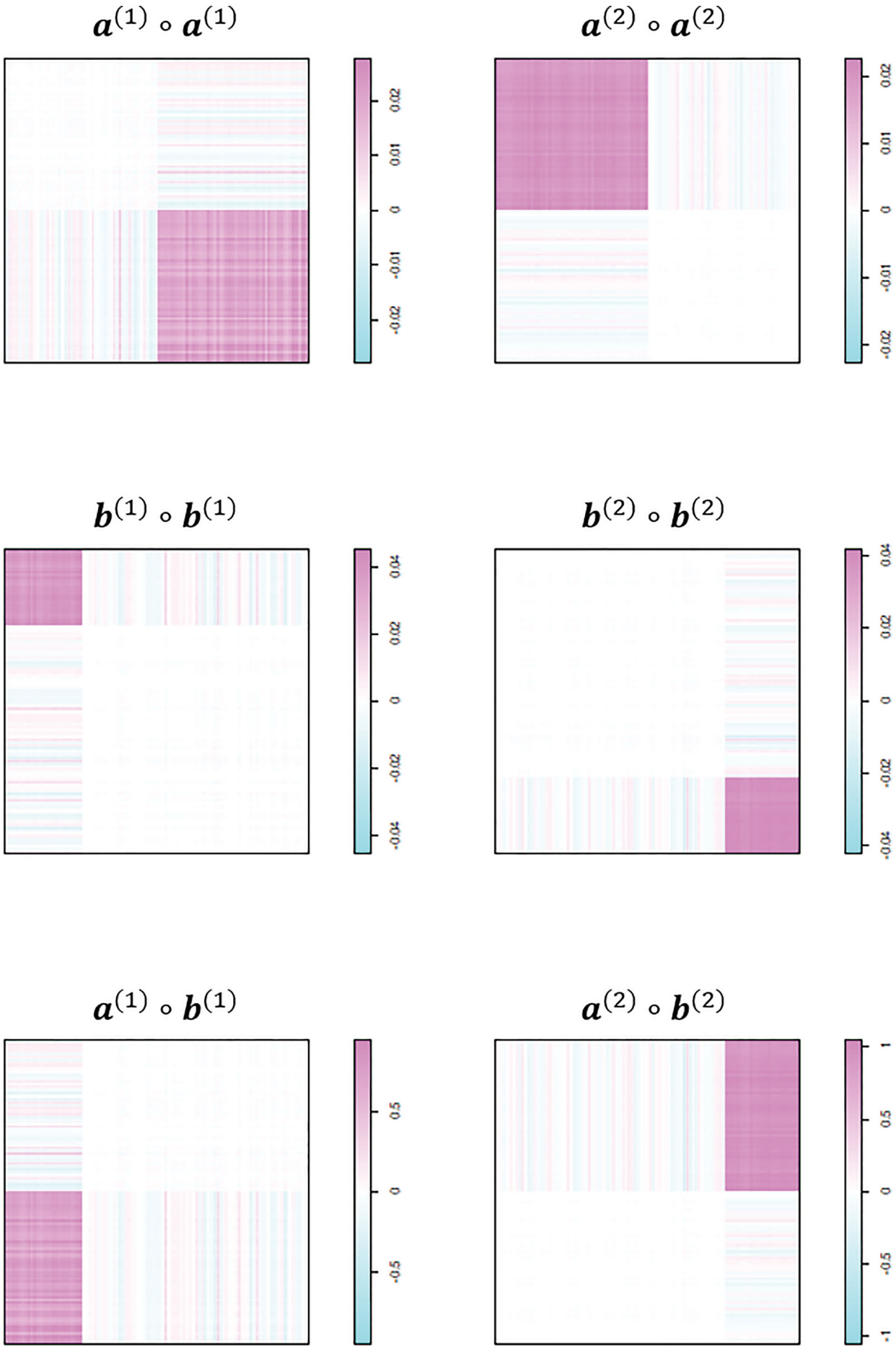

Figure 3. Correlation tensor decomposition from the example of Section 3.1, showing two block-wise correlated regions.

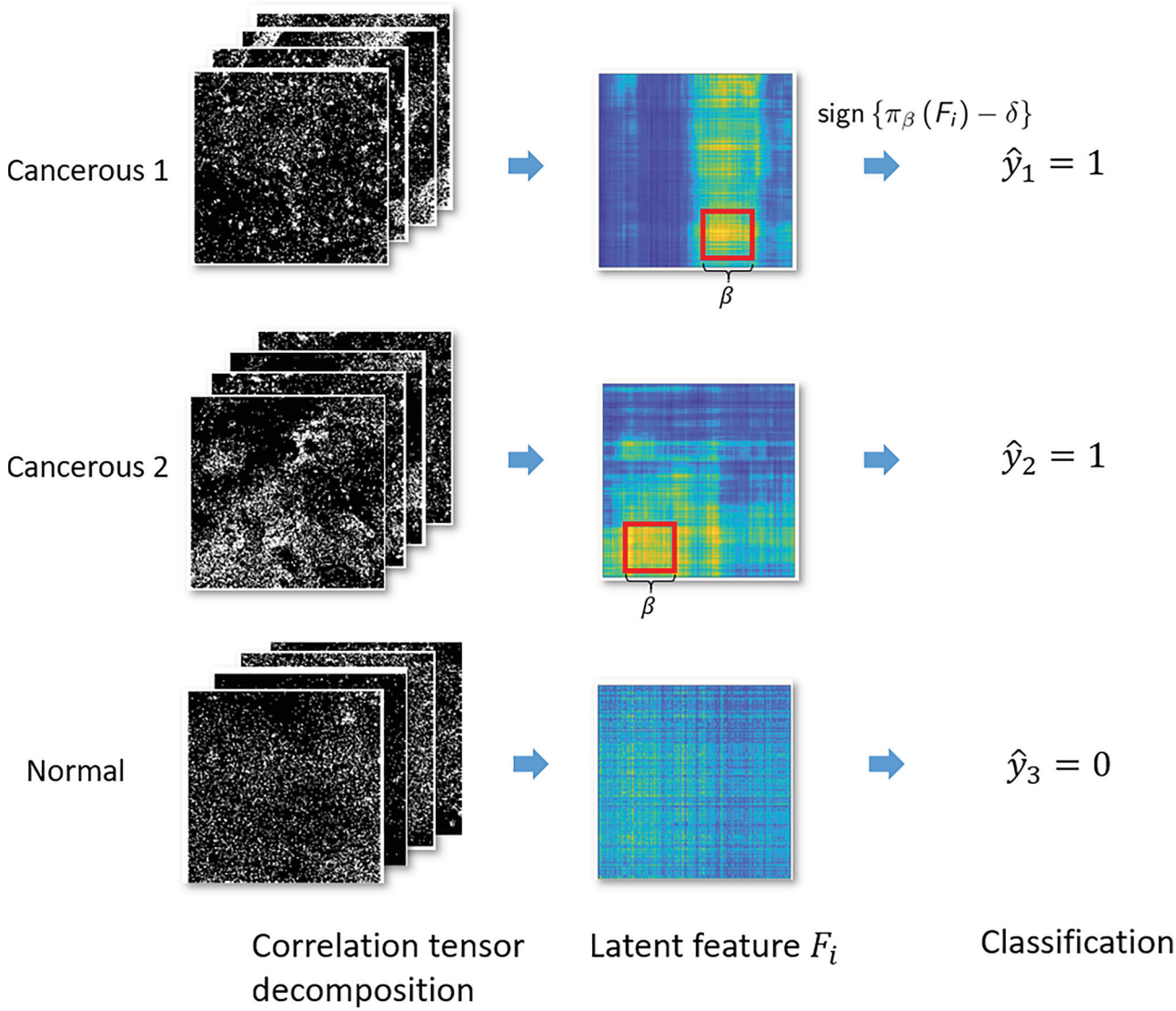

Figure 4. Flowchart of the classification based on the latent features extracted from correlation tensor decomposition.

where $p \leq s$ is due to symmetry, and A, B are defined in Equation (1). The decomposition algorithm updates A and B iteratively. According to Wang et al. (2014), when B is given, the optimal $D^{(ps)}$ can be solved explicitly as

$$D^{(ps)} = \text{Diag}\left[\left\{(B^TB) \odot (B^TB)\right\}^{-1}(B * B)^T \text{vec}(\bar{\mathcal{C}}_{p:s:})\right], \quad (6)$$

where $\odot$ is the element-wise product and $*$ is the Khatari-Rao product. Since the construction of $D^{(ps)}$ utilizes the tubes of $\mathcal{A}$ by fixing the first two dimensions, we can update $\mathcal{A}$ tube by tube, and denote the corresponding estimator as $\hat{\mathcal{A}}$.

For the next step, we estimate $\boldsymbol{a}^{(r)}$ through a rank-1 sparse singular value decomposition (SSVD) of $\hat{\mathcal{A}}_{::r}$, that is,

$$\hat{\boldsymbol{a}}^{(r)} = \arg\min_{\boldsymbol{a}^{(r)}} \|\boldsymbol{a}^{(r)} \circ \boldsymbol{a}^{(r)} - \hat{\mathcal{A}}_{::r}\|_F^2 + \lambda\|\boldsymbol{a}^{(r)}\|_1. \quad (7)$$

The SSVD problem is well-studied and can be solved by the penalized matrix decomposition (PMD) algorithm Witten, Tibshirani, and Hastie (2009), which is described in Algorithm 2. Updating of B with the fixed A follows the same manner.

We remark that the L_p-type penalty also improves the computational stability, since the penalty term can prevent the decomposed components A and B from diverging during the alternative updating process. Specifically, the target tensor involves the parameters A and B in a multiplication form; hence, the scale of A and B could vary significantly without the penalty term, which might lead to computational instability. Therefore, adding a penalty term keeps the scale of A and B relatively consistent and avoids local minimums at the same time. The numerical evidence is included in the supplementary materials.

The above algorithm is efficient in implementation since the updating procedure in Equations (6) and (7) at each step can be solved explicitly. The complete algorithm is summarized in Algorithm 1. The error tolerance ϵ in Algorithm 1 is set as 10^{-3} in the simulation and real data examples. The computational complexity of Algorithm 1 is $\mathcal{O}\{n_{\text{iter}}(c_{\text{HLS}} + Rc_{\text{SSVD}})\}$, where n_{iter} is the number of iterations controlled by the tolerance error, c_{HLS} is the complexity of Steps 2(a) and 3(a) in Algorithm 1, and c_{SSVD} is the complexity of the SSVD (Sun et al. 2017) algorithm for an $L \times L$ matrix. Specifically, Steps 2(a) and 3(a) can be viewed as a generalization of the ordinary least square, with $c_{\text{HLS}} = \mathcal{O}(\text{RL}^4)$, where HLS stands for "high-order least square". In the case without penalty, c_{SSVD} degenerates to the

Algorithm 1 Correlation tensor decomposition algorithm

Input: Sample correlation tensor $\bar{\mathcal{C}}$, rank R, maximum iteration d_{max}, penalty parameter λ and error tolerance ϵ.
Output: Decomposed factor A and B.

1. (*Initialization*) Set $d = 1$, sample initial values from $\mathcal{N}(0,1)$ for $A^{(0)}$ and $B^{(0)}$.
2. (*Given B, update A*) At the dth iteration, fix $B^{(d)}$,
   (a) Compute $\hat{\mathcal{A}}$ with $\hat{\mathcal{A}}_{ps:} = \left\{(B^{(d)T}B^{(d)}) \odot (B^{(d)T}B^{(d)})\right\}^{-1} (B^{(d)} * B^{(d)T})\text{vec}(\bar{\mathcal{C}}_{p:s:})$.
   (b) For each r, update $A_{:r}^{(d+1)}$ with $\hat{\boldsymbol{a}}^{(r)} = \arg\min_{\boldsymbol{a}^{(r)}} \|\boldsymbol{a}^{(r)} \circ \boldsymbol{a}^{(r)} - \hat{\mathcal{A}}_{::r}\|_F^2 + \lambda\|\boldsymbol{a}^{(r)}\|_1$ through Algorithm 2.
3. (*Given A, update B*) Fix $A^{(d+1)}$,
   (a) Compute $\hat{\mathcal{B}}$ with $\hat{\mathcal{B}}_{qt:} = \left\{(A^{(d+1)T}A^{(d+1)}) \odot (A^{(d+1)T}A^{(d+1)})\right\}^{-1} (A^{(d+1)} * A^{(d+1)T})\text{vec}(\bar{\mathcal{C}}_{:q:t})$.
   (b) For each r, update $B_{:r}^{(d+1)}$ with $\hat{\boldsymbol{b}}^{(r)} = \arg\min_{\boldsymbol{b}^{(r)}} \|\boldsymbol{b}^{(r)} \circ \boldsymbol{b}^{(r)} - \hat{\mathcal{B}}_{::r}\|_F^2 + \lambda\|\boldsymbol{b}^{(r)}\|_1$ through Algorithm 2.
4. (*Stopping Criterion*) Calculate the fitted correlation tensor $\hat{\mathcal{C}}^{(d+1)}$ by $\hat{\mathcal{C}}_{stpq}^{(d+1)} = \sum_{r=1}^{R} A_{sr}^{(d+1)} A_{pr}^{(d+1)} B_{tr}^{(d+1)} B_{qr}^{(d+1)}$, and the corresponding loss function $Q^{(d+1)} = L(A^{(d+1)}, B^{(d+1)})$. Stop if $||\hat{\mathcal{C}}^{(d)} - \hat{\mathcal{C}}^{(d+1)}||_F / ||\hat{\mathcal{C}}^{(d)}||_F < \epsilon$ or $|1 - Q^{(d+1)}/Q^{(d)}| < \epsilon$ or $d + 1 > d_{max}$, otherwise set $d \leftarrow d + 1$, and repeat Step 2 and 3.

Algorithm 2 Single-factor penalized matrix decomposition algorithm (PMD) from Witten, Tibshirani, and Hastie (2009)

Input: Positive semidefinite matrix M, penalty parameter λ.
Output: Rank-1 penalized eigenvector $\boldsymbol{v}$.

1. (*Initialization*) Set $\boldsymbol{v}$ to be the unit vector with equal entries. Denote $S(v, \lambda) = \text{sign}(v)(|v| - \lambda)_+$ as the soft thresholding function and $S(\boldsymbol{v}, \lambda) = (S(v_1, \lambda), S(v_2, \lambda), \cdots)$.
2. (*Iterate until convergence*) $\boldsymbol{v} = S(M\boldsymbol{v}, \lambda)/\|S(M\boldsymbol{v}, \lambda)\|_2$.
3. $d \leftarrow \boldsymbol{v}^T M \boldsymbol{v}$, $\boldsymbol{v} \leftarrow \sqrt{d}\boldsymbol{v}$.

complexity of the regular SVD which is $\mathcal{O}(L^3)$, and the total complexity becomes $\mathcal{O}(n_{\text{iter}}\text{RL}^4)$. Note that both Steps 2(a), 3(a) and SSVD can be carried out in parallel to reduce the computation time significantly. In practice, decomposing the correlation tensor from an image of size 100×100 with 10 modalities containing 10^8 elements costs 10 seconds using an Intel Core i7-6700 Processor, 3.4GHz.

In the following, we provide a brief discussion on selecting R in implementation. Indeed, it is always challenging to identify the exact value of the underlying rank R. Many efforts have been made in the recent years to address this problem, for example, BIC-based criteria (Goutte and Amini 2010; Sun and Li 2019), cross-validation (Bro and Kiers 2003; Kolda and Plantenga 2014) and likelihood-based methods (Fu, Matsushima, and Yamanishi 2019). In general, we regard R as a tuning parameter which could be selected through the elbow-point strategy based on the mean square error (MSE) of the recovered tensor compared to the sample correlation tensor. This is similar to selecting the number of principal components in principal component analysis (PCA). We conduct a numerical experiment to illustrate the effectiveness of the elbow-point method. Due to the space limit, we attach the details in the supplementary materials.

Note that the selection of R might not be optimal in practice due to the randomness of observations. However, specifying a larger rank than the true rank would not affect the estimation convergence as it is able to recover the underlying tensor structure completely, although the convergence rate could be affected when M is small since more parameters are involved in estimation (See supplementary materials). In general, we suggest selecting R equal to or slightly larger than the elbow point.

4. Theoretical Result

In this section, we establish theoretical properties regarding the identifiability and the asymptotic theory of the correlation tensor decomposition.

4.1. Identifiability

Although the proposed method does not rely on the identifiable latent factors, as the prediction only depends on the recovered correlation tensor, we provide some brief discussion regarding the identifiability issue in the following. Identifiability is critical in tensor decomposition and could be essential for consequential theoretical development. In the proposed decomposition (1), the identifiability issue is attributed to three parts. The first two are standard indeterminacies of scaling and permutation, and the third one refers to the nonuniqueness of the CP decomposition with more than one possible combinations of rank-one tensors. Specifically, assuming $\boldsymbol{a}^{(r)}$'s and $\boldsymbol{b}^{(r)}$'s are the solution to Equation (1) and their column stacks are denoted as A and B, then the scaling indeterminacy refers to the case that for any diagonal scaling matrices $\Phi = \text{Diag}(\phi_1, \ldots, \phi_R)$, $A\Phi$ and $B\Phi^{-1}$ are also the solutions to Equation (1). The permutation indeterminacy indicates that for any $R \times R$ permutation matrix Π, $A\Pi$, and $B\Pi^{-1}$ are also the solutions.

To deal with the permutation indeterminacy, we rearrange the $\boldsymbol{a}^{(r)}$'s such that

$$\|\boldsymbol{a}^{(1)}\|_2^2 \geq \|\boldsymbol{a}^{(2)}\|_2^2 \geq \cdots \|\boldsymbol{a}^{(R)}\|_2^2,$$

which is equivalent to imposing a descending order based on the vector norm. Then $\boldsymbol{b}^{(r)}$'s can be rearranged using the same order. To deal with the scaling indeterminacy, we can rescale $\boldsymbol{a}^{(r)}$ such that $\boldsymbol{a}^{(r)}(1) = 1$ for $r = 1, \ldots, R$ and rescale $\boldsymbol{b}^{(r)}$ accordingly so that their outer product remains unchanged.

Next, we provide a sufficient condition for the uniqueness of the correlation tensor decomposition.

Proposition 1. Let $\mathcal{C}$ satisfy Equation (1), and A, B be the column stacks of the decomposed components. The decomposition is unique if A and B have full column rank.

The proof is provided in the supplementary materials. Different from the identifiability conditions in Tang, Bi, and Qu (2019), Bi, Qu, and Shen (2018), we consider the semi-symmetric framework of the correlation tensor decomposition

in the proof. In practice, the condition of Proposition 1 is easy to check. In the following proposition, we derive a sufficient condition for the uniqueness of the latent feature defined in Equation (4)

Proposition 2. The latent feature F is unique if A and B have full-column rank.

Notice that there is no indeterminacy of scaling or permutation here since the latent feature is the summation of absolute values of the product between the columns of A and B.

4.2. Asymptotics for Correlation Tensor Estimation

Before we establish our theoretical results, we introduce some notations first. Given M iid random matrix $X^{(m)} \in \mathbb{R}^{L\times L}$ with $\mathbb{E}\left(X^{(m)}\right) = 0_{L\times L}, m = 1,\ldots,M$, the correlation tensor $\mathcal{T}_0$, and covariance tensor $\mathcal{C}_0$, let $\bar{\mathcal{C}}$ be the sample correlation tensor estimator as defined in Equation (3), then $\mathbb{E}\left(\bar{\mathcal{C}}\right) = \mathcal{C}_0$. Let $\Omega = \{\mathcal{C} = (c_{pqst})_{1\le p,q,s,t\le L}: c_{pqst} = \sum_{r=1}^{R} a_p^{(r)} b_q^{(r)} a_s^{(r)} b_t^{(r)}$ for $(p,q) \neq (s,t)\}$ be the parameter space. For any $\mathcal{C} \in \Omega$, we define the loss function regarding the (p,q,s,t)th element of $\bar{\mathcal{C}}$ as

$$\ell(\mathcal{C}, \bar{c}_{pqst}) = (c_{pqst} - \bar{c}_{pqst})^2.$$

Let $J(\mathcal{C})$ be a nonnegative penalty function, then the objective function can be formulated as

$$\text{Loss}(\mathcal{C}|\bar{\mathcal{C}}) = \frac{1}{N_L} \sum_{p,q\neq s,t} \ell(\mathcal{C}, \bar{c}_{pqst}) + \lambda J(\mathcal{C}), \tag{8}$$

where $N_L = L^4 - L^2$ is the number of off-diagonal entries in the correlation tensor.

Let $\Delta(\mathcal{C}, \mathcal{C}_0) = \frac{1}{N}\sum_{p,q\neq s,t}\left\{\ell\left(\mathcal{C}, \bar{c}_{pqst}\right) - \ell\left(\mathcal{C}_0, \bar{c}_{pqst}\right)\right\}$ be the loss difference and $K(\mathcal{C}, \mathcal{C}_0) = \mathbb{E}(\Delta(\mathcal{C}, \mathcal{C}_0))$ be the expected loss difference. It is straightforward that $K(\mathcal{C}, \mathcal{C}_0) \ge 0$ for any $\mathcal{C} \in \Omega$ and $K = 0$ if and only if $\mathcal{C} = \mathcal{C}_0$. Then the distance between $\mathcal{C}$ and $\mathcal{C}_0$ can be defined as

$$d(\mathcal{C}, \mathcal{C}_0) = K^{1/2}(\mathcal{C}, \mathcal{C}_0) = \left\{\frac{1}{N}\sum_{p,q\neq s,t}\left(c_{pqst} - c_{0,pqst}\right)^2\right\}^{\frac{1}{2}}.$$

In practice, the range of correlation coefficients is between -1 and 1. Therefore, it is reasonable to assume that the decomposed factors of $\mathcal{C}$ are bounded so that $\|\boldsymbol{a}^{(r)}\|_\infty, \|\boldsymbol{b}^{(r)}\|_\infty < \eta$, where η is a positive constant.

In our framework, we do not need to impose any distribution assumptions such as the normal distribution on the observations. We only require the following sub-Gaussian property which is commonly adopted in high-dimensional analysis.

Definition 2 (standard sub-Gaussian random vector in $\mathbb{R}^p$). Let Z be a random vector in $\mathbb{R}^p$. Then Z is *standard sub-Gaussian* if there exists an $\tau \ge 0$ such that for all $v \in \mathbb{R}^p$,

$$\mathbb{E}\left(e^{v^T(Z-\mathbb{E}(Z))}\right) \le e^{\tau^2 v^T v/2}.$$

In the following, we provide the convergence rate of the error bound for the proposed estimator. We have a standard regularity condition as follows.

(A1) For any $1 \le p,q,s,t \le L$, $E(X_pX_qX_sX_t) < \infty$.

Condition (A1) assumes a bounded fourth moment. This assumption can be easily validated for imaging data, since the pixel values are restricted to a certain range, for example, $[0, 255]$, so that $X_pX_qX_sX_t$ is always uniformly bounded and thus condition (A1) holds naturally.

Theorem 1. Suppose $\Sigma_0^{-1/2}\text{vec}(X)$ is a standard sub-Gaussian vector, where $\Sigma_0 = \text{corr}\{\text{vec}(X)\}$. Let condition (A1) hold, and $\hat{\mathcal{C}}$ be the minimizer of the loss function (8) with $\lambda = o(R^3M^{-1}L^{-3})$, then for any $0 < u < 1$, we have

$$d\left(\hat{\mathcal{C}}, \mathcal{C}_0\right) \le \omega_{\max}(\Sigma_0)^{3/4}\sqrt{\frac{R^3}{ML^3}\log\frac{1}{u}}, \tag{9}$$

with probability at least $1-u$, for sufficiently large M, L, where $\omega_{\max}(\Sigma_0)$ denotes the largest eigenvalue of Σ_0.

Theorem 1 establishes the asymptotic property for the estimated correlation tensor $\hat{\mathcal{C}}$. First, the recovered correlation structure converges to the true structure as the number of modality M increases with a rate of $\mathcal{O}(M^{-\frac{1}{2}})$, which is consistent with the rate of the sample correlation estimator. Furthermore, as the tensor size L increases, Theorem 1 provides an extra convergence rate of $L^{-\frac{3}{2}}$ if $\omega_{\max}(\Sigma_0)$ is bounded, which essentially benefits from the proposed low-rank tensor decomposition model, yielding a substantially smaller parameter space compared with the increasing data volume. Note that $L^{-\frac{3}{2}}$ is the optimal rate for the CP-type tensor decomposition in the literature (Wang and Li 2020). In addition, we allow the rank R to grow with a rate smaller than $M^{\frac{1}{3}}L$ while the convergence is still ensured.

On the other hand, the underlying correlation also plays an important role in the error bound. Indeed, this is one of the major challenges encountered in theoretical analysis for correlation tensor estimation. We quantify the effect of correlation on the convergence rate by $\omega_{\max}(\Sigma_0)$, which is introduced by the generalized Hanson-Wright inequality in $\mathbb{R}^p$ Chen et al. (2021). Note that $\omega_{\max}(\Sigma_0)$ may grow with L. In the following Corollary 1, we provide the error bound under regularity conditions imposed on $\mathcal{C}_0$, which contains a wide class of correlation structures.

Corollary 1. Assume that the conditions for Theorem 1 hold, then for any $0 < u < 1$, we have

(Case I) if $c_{0,pqst} = \rho > 0$ for all $(p,q) \neq (s,t)$, then

$$d\left(\hat{\mathcal{C}}, \mathcal{C}_0\right) \le 2\rho^{\frac{3}{4}}R^{\frac{3}{2}}M^{-\frac{1}{2}}\sqrt{\log(1/u)};$$

(Case II) if $c_{0,pqst} < \rho^{|p-s|+|q-t|}$, where $0 < \rho < 1$, then

$$d\left(\hat{\mathcal{C}}, \mathcal{C}_0\right) \le \left(1 + 4\rho + \frac{8\rho^2}{(1-\rho)^2}\right)R^{\frac{3}{2}}M^{-\frac{1}{2}}L^{-\frac{3}{2}}\sqrt{\log(1/u)},$$

with probability at least $1-u$.

Case (I) describes a strong correlation pattern, referring to an exchangeable correlation structure, where the correlation

coefficients are the same between any two pixels. Under this setting, due to the substantially high correlation among the data, there would be convergence only with an increasing number of modalities M. In Case (II), the correlation between two pixels X_{pq} and X_{st} decays geometrically, which implies a weaker correlation among the pixels that are far away from each other. Many commonly adopted correlation structures under the spatial–temporal data framework fall into Case (II); for example, the Gaussian model and the Matérn model (Wackernagel 2003; Gaetan and Guyon 2010; Chiles and Delfiner 2009; Schlather 1999) in geostatistics and Lattice models (Lampert, Ralaivola, and Zimin 2018) in spatial analyses. Another special case is when the X_{pq}'s are all uncorrelated or independent. Corollary 1 indicates that the convergence rate under Case (II) will attain the optimal rate as $M^{-1/2}L^{-3/2}$. Furthermore, note that $\mathcal{O}(1) \leq \omega_{\max}(\Sigma_0) \leq \mathcal{O}(L^2)$ holds for an arbitrary correlation structure. Hence, as long as $\omega_{\max}(\Sigma_o) = o(L^2)$, our error bound will decrease as the tensor size L increases, which is advantageous compared to the sample correlation estimate.

We remark that Theorem 1 provides the fundamental estimation consistency properties of the correlation tensor decomposition with a general class of penalized models, where we only assume that $\mathcal{C}_0$ is low-rank. Moreover, the error bound can be further improved by imposing appropriate penalty functions according to additional prior knowledge on the structure of $\mathcal{C}_0$. For example, an L_1 penalty is commonly used to encourage sparsity of the coefficients and has been intensively studied in various settings (Meinshausen and Yu 2009; Negahban et al. 2012; Raskutti, Yuan, and Chen 2019; Zhang and Han 2019). In the following, we provide a tighter bound by imposing the L_1 penalty on decomposition components A and B, when the underlying correlation is assumed sparse.

We denote the index set of nonzero entries in $\boldsymbol{a}^{(r)}$ and $\boldsymbol{b}^{(r)}$ of $\mathcal{C}_0$ as $\mathcal{I}_a^r$ and $\mathcal{I}_b^r$, respectively, and let $k_a = |\cup_r \mathcal{I}_a^r|$ and $k_b = |\cup_r \mathcal{I}_b^r|$, where $|\cdot|$ denotes the set cardinality.

Corollary 2. Let $\hat{\mathcal{C}}$ be the minimizer of the loss function in Equation (8) with penalty $J(\mathcal{C}) = \sum_{r=1}^R |\boldsymbol{a}^{(r)}|_1 + \sum_{r=1}^R |\boldsymbol{b}^{(r)}|_1$. Suppose all conditions in Theorem 1 hold and $\lambda = \mathcal{O}(R^3M^{-1}L^{-3})$, then for any $0 < u < 1$, we have

$$d(\hat{\mathcal{C}}, \mathcal{C}_0) \leq \left[1 + \frac{k_a k_b}{L^2}\{\omega_{\max}(\Sigma_0) - 1\}\right]^{\frac{3}{4}} \sqrt{\frac{R^3}{ML^3}\log\frac{1}{u}}, \quad (10)$$

with probability at least $1 - u$, for sufficiently large M and L.

Note that $k_a k_b$ quantifies the signal area of the 2D image, where $k_a \leq L$ and $k_b \leq L$. When the pixels are strongly correlated, $\omega_{\max}(\Sigma_0)$ may grow as L increases. In this case, Corollary 2 indicates that the error bound would be further improved under the sparsity setting given $k_a \ll L$ and $k_b \ll L$, compared to Theorem 1. This is analogous to the results in Zhang and Han (2019), which considers a sparse SVD based on thresholding, but requiring independent tensor entries.

4.3. Consistency of Classification

In addition to the recovery of the correlation tensor structure, we are also interested in identifying the signal area through the latent features defined in Equation (4) as they contribute to the classification for disease diagnosis.

Formally, we define the *signal areas* of an image as the index set $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_{n_\mathcal{S}}$, where for any $(s,t) \in \mathcal{S}_j, j = 1, \ldots, n_\mathcal{S}$, $\text{corr}(X_{pq}, X_{st}) > 0$. Note that we allow more than one signal area in cancerous subject, that is, $n_\mathcal{S} \geq 1$, where each signal area is a collection of mutually correlated pixels. For simplicity, we restrict $n_\mathcal{S} = 1$ in the following discussion and use only $\mathcal{S}$ to denote the signal area. Cases of $n_\mathcal{S} > 1$ can be generalized similarly.

In traditional image analyses, the signals are identified based on the magnitude of pixels, while the stronger correlation could decrease the convergence rate (Tang, Bi, and Qu 2019). In contrast, in our model the signal areas are defined as the pixels which are highly correlated with each other. The larger correlation can improve the classification performance by enhancing the margin between zero and nonzero correlations, which is supported by the following theorem.

Theorem 2. Let F be the latent feature defined in Equation (4), and denote $\hat{F}$ as its estimator. Assume the conditions in Theorem 1 hold, then for any $0 < \epsilon < \bar{\rho}_0$,

$$P\left(\|\hat{F}\|_\mathcal{S}^2 > \epsilon\right) \geq 1 - 6\exp\left\{-\frac{1}{2}M\frac{|\mathcal{S}|^{\frac{3}{2}}(\bar{\rho}_0 - \epsilon)^2}{R^3\|\Sigma_0\|_2^{\frac{3}{2}}}\right\}, \quad (11)$$

where $\|F\|_\mathcal{S}^2 = \frac{1}{|\mathcal{S}|}\sum_{(p,q)\in\mathcal{S}} f_{pq}^2$, and $\bar{\rho}_0 = \sqrt{\sum_{(p,q),(s,t)\in\mathcal{S}} c_{0,pqst}^2}$.

Theorem 2 bridges the gap between the estimated correlation tensor and the recovered latent feature in classification. Theorem 2 indicates that as M grows, the average F-norm of $\mathcal{S}$ is greater than ϵ with probability tending to 1. The tail probability in (11) is also influenced by the average correlation coefficient $\bar{\rho}_0$. Since the magnitude of $\|\Sigma_0\|_2$ is at most the same as $|\mathcal{S}|$, increasing $\bar{\rho}_0$ also reduces the tail probability, and thus the proposed method is more powerful in distinguishing the signal region from the background region.

Consequently, we prove the classification consistency using the estimated latent feature. Let β_0 and δ_0 be the two thresholding parameters which determine the label of the image as in Equation (5). Furthermore, we denote $\mathcal{R}(\hat{\mathcal{L}}) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(y_i, \hat{F}_i, \hat{\beta}, \hat{\delta})$ as the empirical risk for the sample loss function, and $\mathcal{R}(\tilde{\mathcal{L}}) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(y_i, \hat{F}_i, \beta_0, \delta_0)$ as the empirical risk of the oracle classifier. The next result provides the consistency of the empirical risk.

Corollary 3. Assume that the conditions of Theorem 1 hold, then as M and $n \to \infty$, we have

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

Corollary 3 implies that as the number of subjects and the number of image modalities of each subject grows, the empirical

risk of the sample loss with the learned parameter converges to the empirical risk with the oracle classifier.

Moreover, by directly applying Theorem 2, we show that the sensitivity of the proposed method satisfies

$$P\left\{\pi_\beta(\hat{F}) > \delta \mid Y = 1\right\} \geq 1 - 6 \times \exp\left\{-\frac{1}{2}M \min\left(\frac{|\mathcal{S}|^2}{\|\Sigma_0\|_F^2}\left(\bar{\rho}_0 - \frac{\beta^2\delta}{|\mathcal{S}|}\right)^2, \frac{|\mathcal{S}|}{\|\Sigma_0\|_2}\left(\bar{\rho}_0 - \frac{\beta^2\delta}{|\mathcal{S}|}\right)\right)\right\}. \quad (12)$$

This implies that if the signal area is more concentrated, that is, $\frac{|\mathcal{S}|}{\beta_0^2}$ is larger, or the average correlation within the signal area $\bar{\rho}_0$ is bigger, the sensitivity is higher.

The detailed proofs of the theoretical properties in this section can be found in the supplementary materials.

5. Simulation Study

5.1. Simulation 1: Estimation Efficiency

In this subsection, we illustrate the estimation efficiency of the correlation tensor with the proposed method. We consider a series of images of size $L \times L$ with M modalities. Each $X^{(m)}$ is generated from a normal distribution with mean 0 and a correlation tensor with the primal diagonal components $\text{var}(X^{(m)}_{p,q}) = 1$, and the off-diagonal components following $\mathcal{C} = \sum_{r=1}^{2} \boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)} \circ \boldsymbol{a}^{(r)} \circ \boldsymbol{b}^{(r)}$, where

$$\begin{aligned}
\boldsymbol{a}^{(1)} &= \sqrt{0.9}(\underbrace{1, \ldots, 1}_{L/2}, 0, \ldots, 0),\\
\boldsymbol{b}^{(1)} &= \sqrt{0.8}(\underbrace{1, \ldots, 1}_{L/4}, 0, \ldots, 0),\\
\boldsymbol{a}^{(2)} &= \sqrt{0.9}(0, \ldots, 0, \underbrace{1, \ldots, 1}_{L/2}),\\
\boldsymbol{b}^{(2)} &= \sqrt{0.8}(0, \ldots, 0, \underbrace{1, \ldots, 1}_{L/4}).
\end{aligned}$$

Under this setting, the highly correlated pixels of X are located in two block regions similar to the example in Section 3.1.

We compare the performance of the proposed method with the sample correlation estimator (3) and the principal orthogonal complement thresholding method (POET) (Fan, Liao, and Mincheva 2013). The estimation efficiency is evaluated by the average mean square error (MSE) of $\mathcal{C}$ based on 100 replications. In particular, we investigate the asymptotics of MSE under two settings: (a) increasing M, with fixed $L = 40$; (b) increasing L, with fixed $M = 10$. Tables 1 and 2 summarize the results of setting (a) and (b), respectively. We also provide corresponding line plots in Figure 5 to illustrate the discrepancy of different methods more intuitively. The results show that the proposed method achieves the lowest MSE in each setting, which reduces the MSE by at least 90% compared with the other two methods. Moreover, the proposed method benefits from the increase of the image size L in addition to the number of modalities M, while the MSE of the sample covariance method and POET only decrease as M increases. This is because the sample covariance method only uses pairwise pixel information, while the proposed method borrows information from neighboring pixels based on the low-rank tensor structure. Although the POET estimator considers low-rank structure as well, the corresponding MSE is always higher than the proposed method since it only imposes a one-dimensional factor model without fully using the spatial structure of the imaging data. In addition, the discrepancy between POET and the sample covariance approach is not obvious when M is small, while the proposed method performs consistently better regardless of M.

Table 1. Estimation accuracy of Simulation 1 with various M: average MSE based on 100 replications with standard deviation.

M	10	20	30	40	50
Sample covariance	1.13(0.14)	0.61(0.04)	0.43(0.02)	0.33(0.02)	0.27(0.01)
POET	0.73(0.14)	0.27(0.04)	0.15(0.02)	0.10(0.02)	0.07(0.01)
Proposed	**0.02(0.01)**	**0.01(0.00)**	**0.01(0.00)**	**0.01(0.00)**	**0.01(0.00)**

Table 2. Estimation accuracy of Simulation 1 with various L: average MSE based on 100 replications with standard deviation.

L	10	20	30	40	50
Sample covariance	1.11(0.23)	1.13(0.16)	1.14(0.15)	1.13(0.14)	1.12(0.14)
POET	0.80(0.23)	0.75(0.17)	0.73(0.14)	0.73(0.14)	0.73(0.14)
Proposed	**0.10(0.03)**	**0.04(0.01)**	**0.02(0.01)**	**0.02(0.01)**	**0.02(0.00)**

5.2. Simulation 2: Multi-Modality Image Classification

In this simulation, we investigate the prediction performance with multi-modality imaging data which mimics the microvesicle imaging patterns of early-stage breast cancer. For the mth modality of a single subject, the observed data $X^{(i,m)} \in \mathbb{R}^{L \times L}$ is composed of three parts:

$$X^{(i,m)} = S^{(i,m)} + F^{(i,m)} + E^{(i,m)}, \quad m = 1, \ldots, M,\ i = 1, \ldots, N,$$

where $S^{(i,m)}$ represents signal patterns associated with spatially highly correlated pixels, $F^{(i,m)}$ represents the correlated "signal-like" noise without specific spatial patterns, and $E^{(m)}$ represents the noise background.

Specifically, the location of the highly correlated pixels of $F^{(m)}$ follows a Poisson point process with the intensity of a signal $\nu = 75$, and the corresponding intensity of the pixels follows a multivariate normal distribution $\text{MVN}_{n_F}(0, 0.9\mathbf{1}\mathbf{1}^T + 0.1\mathbf{I})$ with exchangeable correlations, where n_F is the number of "signal-like" noise. The noise in $E^{(i,m)}$ is generated from a standard normal distribution.

In this simulation, the label of the subject is based on the pattern of $S^{(m)}$. For a normal subject i, we let $S^{(i,m)} = 0$ for $m = 1, \ldots, M$ and label $h_i = 0$, otherwise the subject is labeled as cancerous with $y_i = 1$. The signal pixels are generated based on the following two settings:

1. *Random blocks*: Similar pattern as described in Simulation 1, where

$$\begin{aligned}
\boldsymbol{a}^{(r)} &= \sqrt{0.9}(0, \ldots, 0, \underbrace{1, \ldots, 1}_{L_b}, 0, \ldots, 0),\\
\boldsymbol{b}^{(r)} &= \sqrt{0.8}(0, \ldots, 0, \underbrace{1, \ldots, 1}_{L_b}, 0, \ldots, 0),
\end{aligned}$$

where $L_b = 0.1L$.

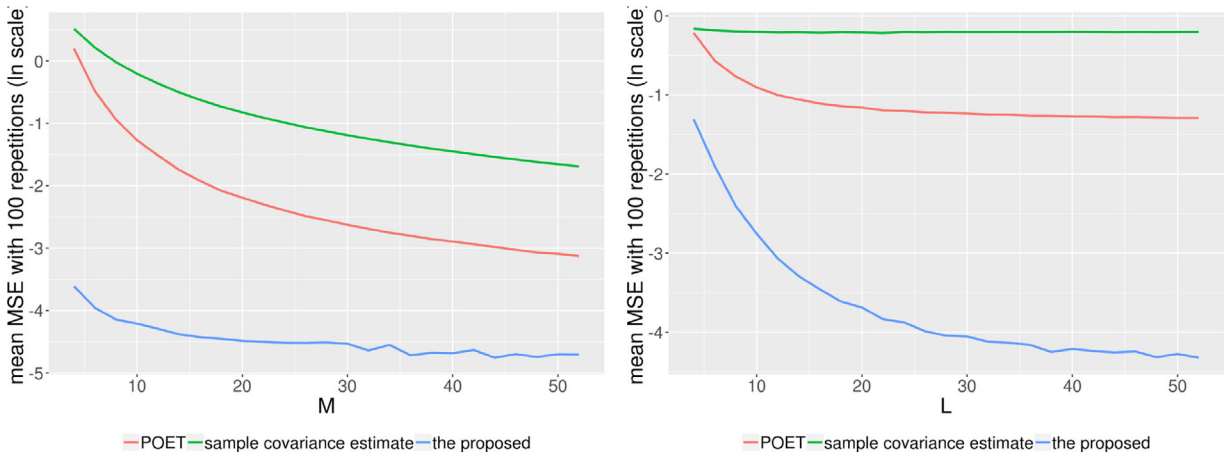

Figure 5. Estimation efficiency comparison with two random block-wise highly correlated pixel areas in Simulation 1. *Left*: MSE change with $L = 40$ and varying M. *Right*: MSE change with $M = 10$ and varying L.

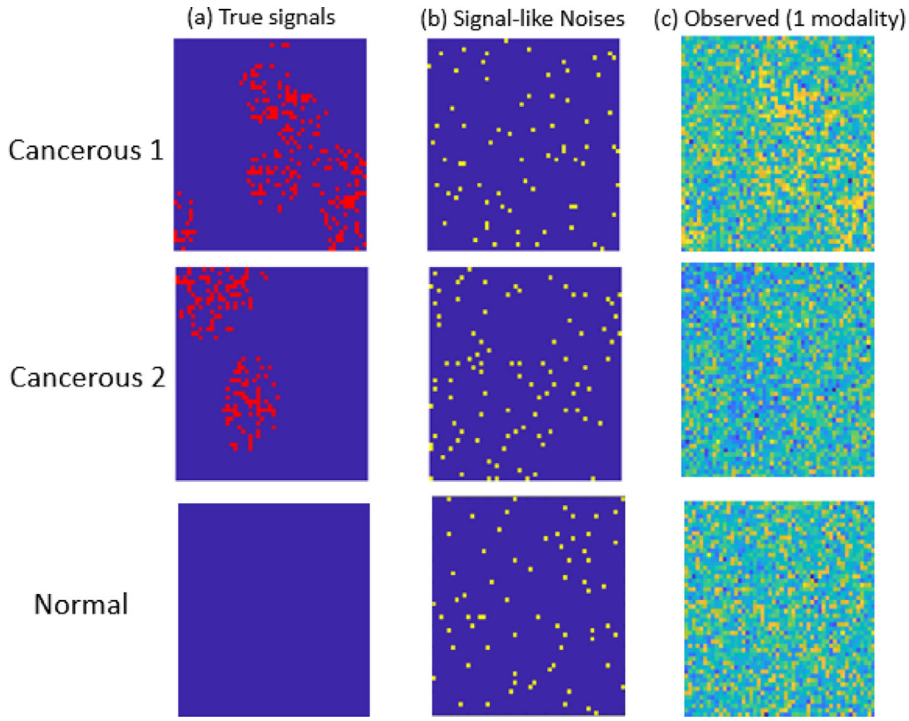

Figure 6. Selected samples with 1-modality in Simulation 2: (a) True signal location. (b) Signal-like noise location. (c) Observed samples.

2. *Random clusters*: The location for the signal pixels follows a Matérn cluster process (Matérn 2013) with intensity of the cluster center $\kappa = 4$, and the mean number of each cluster $\mu = 50$.

In the *random block* setting, L_b is the size of the signal block. That is, there are two L_b by L_b block regions of correlated pixels and the locations of the signal blocks is random. In the *random cluster* setting, the signal regions consist of the clusters with a fixed radius, and the location of the centers of the clusters are random, where κ and μ control the average number of clusters and the density of the signal pixels, respectively. The signal strength of both settings follows $\text{MVN}_{n_S}(0, 0.9\mathbf{1}\mathbf{1}^T + 0.1\mathbf{I})$ where the dimension n_S is the number of the signal pixels. Figure 6 illustrates the composition of the simulated data of both normal subjects and cancerous subjects under the random block setting.

For both of the settings, we let $M = 10$, and generate training, validation and testing sets of size 100, 40, and 60, respectively, with 50% of healthy subjects and 50% of cancer subjects for each set. We let $L = 100$, 200, and 500, and compare the proposed classification method described in Section 3.2 with the higher order CP decomposition method (HOCPD) Tang, Bi, and Qu (2019) which used the components estimated from the traditional CP-decomposition as the input to a logistic regression model, the marginal principal component analysis (MPCA) by Caffo et al. (2010), the vectorizing L_1-penalized logistic regression model (VPL), the tensor regression (TR) model (Zhou, Li, and Zhu 2013) and the convolutional neural network (CNN).

For HOCPD and MPCA, we use the validation set to select the best number of components. For CNN, we use the Python library `Keras` (Chollet 2015) to train the model and

adopt `scikit-optimize` library (*https://github.com/scikit-optimize/scikit-optimize/tree/master/skopt*) to tune the hyperparameters. The computation details can be found in supplementary materials.

The classification results are summarized in Tables 3 and 4, which show that the proposed method outperforms the other methods in terms of average accuracy, sensitivity and specificity on detection of signal block $S^{(i,m)}$. The accuracy of other methods is around 50% due to weak marginal signals, while the proposed method is able to achieve an accuracy of 90% when L is large. Moreover, the proposed method achieves higher classification accuracy under both random block and random cluster settings when L increases, which supports the conclusion of Equation (12) empirically. In practice, it implies that leveraging the resolution of the image could enhance signal identification using the proposed method. In addition, although the random cluster setting does not satisfy the block signal assumption on the spatial correlation pattern, the proposed method still achieves high accuracy, indicating that the proposed method is quite robust as long as the target signals are concentrated in a spatial region. On the other hand, the competing methods only use the marginal intensity information and fail to fully utilize the spatial correlation and correlation information across different modalities.

To illustrate the signal regions identified by the proposed method, Figure 7 provides the identified latent features defined in (4), which clearly shows that the regions of true signals are successfully captured while the randomly scattered signal-like noises are not selected.

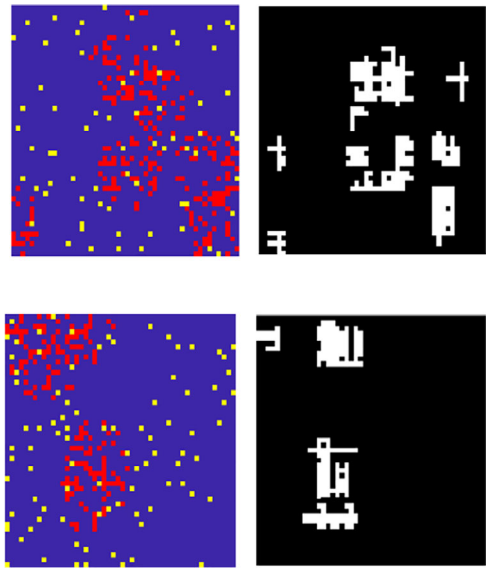

Figure 7. Detected regions of two samples. *Color image*: location of true signals (red) and signal-like noise (yellow); *Black and white image*: latent features identified by the proposed method.

6. Multiphoton Imaging Data Classification

We apply the proposed method to multimodal breast cancer imaging data Tu et al. (2016) provided by Boppart's biophotonics imaging lab in the University of Illinois at Urbana Champaign. There are four modalities for each image: two-photon auto-fluorescence (2PAF), three-photon auto-fluorescence (3PAF), second-harmonic generation (SHG) and third-harmonic generation (THG). These co-localized images are collected based on different contrasts in the micro-environment of the breast tissue at different molecular levels. Figure 8 shows two regions that contain spatially concentrated TMVs in red circles. The signal strength is strong in the 2PAF and 3PAF modalities yet is relatively weak in the other two modalities. We apply the proposed method to integrate information from all modalities to detect TMVs effectively.

To better preserve the small-scale TMVs, we filter out the irrelevant background imaging. Specifically, we preprocess the images to remove the modality-specific background by applying the Gaussian filter MATLAB Image Process Toolbox (2018a) and we also subtract the mean of pixels across four modalities. An illustration of Gaussian filter can be found in Figure 5 of

Table 3. Classification results of Simulation 2 under random block setting with various size of L: average accuracy, sensitivity and specificity based on 100 replications with standard deviations.

L	100			200			500		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
HOCPD	0.51(0.07)	0.51(0.13)	0.51(0.14)	0.50(0.07)	0.48(0.13)	0.52(0.16)	0.50(0.05)	0.51(0.16)	0.49(0.18)
MPCA	0.51(0.07)	0.48(0.10)	0.54(0.10)	0.52(0.06)	0.39(0.09)	0.65(0.08)	0.51(0.07)	0.39(0.09)	0.63(0.10)
VPL	0.50(0.04)	**0.82(0.24)**	0.17(0.24)	0.49(0.04)	**0.83(0.25)**	0.16(0.23)	0.50(0.04)	0.82(0.24)	0.19(0.26)
TR	0.49(0.08)	0.50(0.12)	0.48(0.13)	0.51(0.05)	0.45(0.34)	0.57(0.33)	0.49(0.06)	0.50(0.24)	0.48(0.24)
CNN	0.56(0.13)	0.47(0.21)	0.64(0.24)	0.57(0.16)	0.47(0.23)	0.68(0.26)	0.50(0.05)	0.43(0.33)	0.58(0.34)
Proposed	**0.75(0.05)**	0.58(0.13)	**0.87(0.11)**	**0.84(0.05)**	0.73(0.07)	**0.96(0.05)**	**0.95 (0.02)**	**0.92 (0.05)**	**0.99 (0.02)**

Table 4. Classification results of Simulation 2 under random cluster setting with various size of L: showing average accuracy, sensitivity and specificity based on 100 replications with standard deviations.

L	100			200			500		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
HOCPD	0.47(0.06)	0.47(0.13)	0.48(0.12)	0.49(0.06)	0.49(0.11)	0.48(0.12)	0.55(0.09)	0.47(0.13)	0.64(0.22)
MPCA	0.46(0.06)	0.46(0.10)	0.46(0.10)	0.45(0.06)	0.46(0.09)	0.44(0.10)	0.64(0.08)	0.46(0.10)	0.83(0.14)
VPL	0.50(0.01)	**0.98(0.09)**	0.01(0.07)	0.50(0.00)	**0.99(0.06)**	0.01(0.05)	0.50(0.04)	0.81(0.25)	0.19(0.25)
TR	0.50(0.08)	0.50(0.13)	0.50(0.11)	0.50(0.04)	0.47(0.37)	0.53(0.37)	0.49(0.06)	0.47(0.24)	0.51(0.24)
CNN	0.56(0.10)	0.40(0.17)	0.72(0.22)	0.61(0.10)	0.44(0.19)	0.78(0.23)	0.52(0.07)	0.45(0.31)	0.58(0.33)
Proposed	**0.90 (0.04)**	0.82 (0.06)	**0.97 (0.07)**	**0.98 (0.02)**	0.97 (0.03)	**0.98 (0.02)**	**0.99 (0.02)**	**0.99 (0.01)**	**0.99 (0.02)**

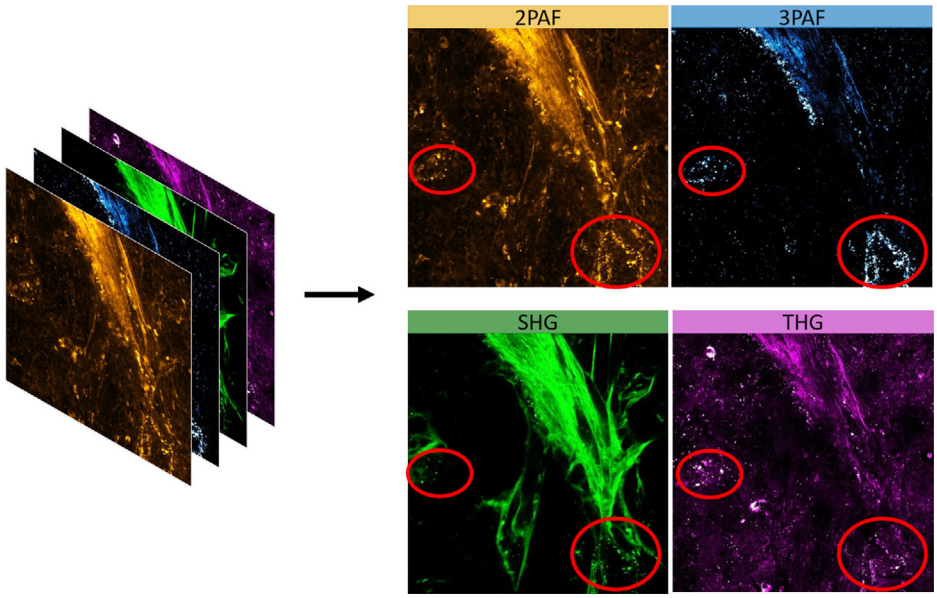

Figure 8. Multiphoton image of a cancer subject with four modalities. TMVs (bright dots in red circle) may have different signal intensities on different modalities.

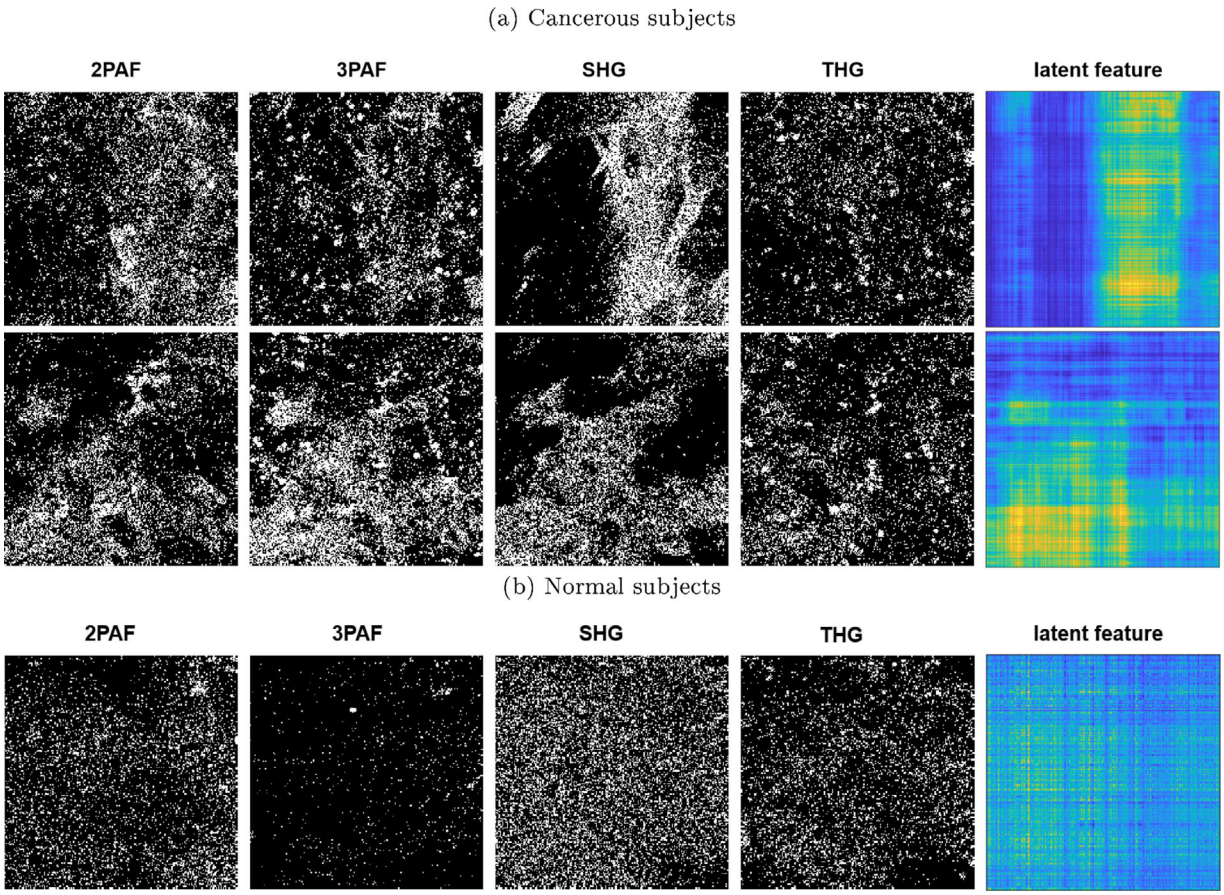

Figure 9. The first four columns are four-modality imaging data of two cancerous subjects and two normal subjects after filtering and segmentation. The last column shows the latent feature detected by the correlation tensor decomposition. The contrast of the images has been adjusted for illustration purpose.

supplementary materials. Note that the preprocessing step does not guarantee removal of all the irrelevant imaging patterns. However, we can treat the remaining noninformative pixels as random noise.

Furthermore, a previous study Tu et al. (2016) showed that the informative TMVs are frequently observed in the microenvironment between certain cellular tissues such as at the lipid boundary and around the stromal regions. Thus, we segment the filtered images into 200 × 200 pixels, and mainly focus on imaging within the potential target locations. Consequently, each sample image is a 200 × 200 × 4 tensor. The left four columns of Figure 9 illustrate the four modalities of the imaging

Table 5. Classification result of breast cancer imaging data based on 100 replications with training size=20, validation size=20 and testing size=20.

Method	Accuracy	Sensitivity	Specificity
HOCPD	0.592(0.108)	0.433(0.180)	0.750(0.184)
MPCA	0.646(0.115)	0.564(0.195)	0.728(0.168)
VPL	0.514(0.066)	0.328(0.358)	0.700(0.341)
TR	0.505(0.067)	0.478(0.446)	0.532(0.454)
CNN	0.753(0.111)	0.733(0.176)	0.773(0.226)
The proposed	**0.814(0.081)**	**0.849(0.123)**	**0.780(0.165)**

data of two cancerous subjects and one normal subject after preprocessing.

We split the preprocessed data into a training set, a validation set and a testing set. Each set consists of 10 samples from normal subjects and 10 from cancerous subjects. We compare the proposed method with the other four methods described in Section 5. The rank of HOCPD and MPCA is determined by a validation set and the architecture of the CNN is tuned in the same way as in Section 5.2. We evaluate the performance of each method by the prediction accuracy, sensitivity and specificity on a testing dataset based on 100 replications, and summarize the results in Table 5, which indicates that the proposed method outperforms the other methods. Specifically, the proposed method improves on the average prediction accuracy by 32%, 30%, 55%, and 5% compared to HOCP, MPCA, VPL, and CNN, respectively. We also notice that the proposed method achieves much higher sensitivity and gives us more power to correctly detect the cancer risk. Achieving high sensitivity is crucial in the early diagnosis of breast cancer, especially when the proportion of the potential cancer patients is small compared with the general population.

The VPL method performs poorly with an average prediction accuracy of 51%. This is probably due to the fact that vectorization is not efficient capturing the spatial relationships among pixels. The HOCPD and MPCA perform better than the VPL, but still suffer low sensitivity due to weak marginal signal intensity from the TMVs and the randomness of signal locations. The CNN provides an acceptable prediction accuracy of 75.3%, yet the sensitivity is still lower than the proposed method, likely because of the high-intensity noise background. In addition, the CNN is not robust against a small sample size and heavily relies on the hyperparameter tunings.

More importantly, the CNN is not able to provide interpretable results. In contrast, the proposed method provides sensible interpretation through latent features. Specifically, the rightmost columns of Figure 9 show heatmaps of latent features captured by the correlation tensor decomposition which are consistent with the observed patterns of TMV signals. The highlighted blocks represent the highly correlated pixels and are extracted features for TMV classification. Comparing the cancerous subjects and the normal subject, it is clear that the latent features of cancerous subjects indicate a more dense pattern of TMV signals while those of the normal subjects are more randomly and sparsely scattered. In particular, the highlighted latent features of the topmost graph from a cancerous subject indicate a vertical area which is consistent with the pattern of the TMV signals in the SHG modality, and the yellow latent features of the top graph from a cancerous subject at the lower-left region of images indicate that these pixels are highly correlated across the modalities. In summary, the proposed method provides an interpretable visualization of the detected TMV signals, and the latent features can serve as a prognostic tool for cancer diagnosis.

7. Discussion

In this article, we introduce the concept of correlation tensor and propose a semi-symmetric tensor decomposition to achieve high estimation accuracy when the size of image is large, and the number of modality is limited. The key idea is to extract the block-wise spatially correlated pixels and informatively reduce the dimension of parameters. In addition, we develop a classification method based on the extracted latent features.

A major contribution of the proposed method is that we are able to preserve the spatial information through correlation tensor decomposition. This facilitates the detection of TMVs where the target signals are both highly correlated and spatially concentrated. Moreover, our numerical and theoretical analyses show that increasing imaging resolution improves signal detection efficiency, and thus benefits classification even with a limited number of modalities.

The proposed decomposition method is able to provide meaningful interpretation. The latent features constructed from the decomposed components provide the locations of block-wise correlated pixels, which is more advantageous compared to Tang, Bi, and Qu (2019), and are useful in the medical imaging diagnosis. In addition, the proposed method can be applied in spatial analysis with repeated observations, especially when the target signals are determined by their correlations, such as in longitudinal fMRI data analysis (O'Brien et al. 2010), remote sensing data (Li et al. 2018b) and calcium imaging data (Soltanian-Zadeh et al. 2019). The comprehensive analyses of these types of data require computation of correlation pairs between pixels or voxels, which is challenging due to the data size. The proposed correlation tensor decomposition method can improve the estimation efficiency as well as enhancing signal detection.

In our model, we detect the correlated pixels by recovering block-wise correlated regions. Although in practice the shape of the signal area might not be rectangular, our proposed method is still able to capture the main signals through a low rank approximation successfully, which is verified through the numerical experiment. Furthermore, we can generalize the current model to imaging data beyond two dimensions such as voxel data from the fMRI, where the dimension of the correlation tensor could be more than four. These are worth further investigation as future research directions.

Supplementary Materials

The online supplement contains the generalized model with cross-modality correlations, technical proofs, additional numerical results and computational details for CNN.

Acknowledgments

The authors are grateful to reviewers, the associate editor and editor for their insightful comments and suggestions which have improved the article significantly. The authors also appreciate Biophotonics Imaging Laboratory

of University of Illinois, Urbana-Champaign for providing the breast cancer image data.

Funding

The work is supported by NSF (grants DMS 1952406 and DMS 1821198).

References

Allen, G. (2012), "Sparse Higher-Order Principal Components Analysis," in *Artificial Intelligence and Statistics*, pp. 27–36. [2,4]

Bar-Joseph, Z., Gifford, D. K., and Jaakkola, T. S. (2001), "Fast Optimal Leaf Ordering for Hierarchical Clustering," *Bioinformatics*, 17, S22–S29. [2]

Bi, X., Qu, A., and Shen, X. (2018), "Multilayer Tensor Factorization With Applications to Recommender Systems," *The Annals of Statistics*, 46, 3308–3333. [2,8]

Bowman, F. D., Guo, Y., and Derado, G. (2007), "Statistical Approaches to Functional Neuroimaging Data," *Neuroimaging Clinics of North America*, 17, 441–458. [1]

Bro, R., and Kiers, H. A. (2003), "A New Efficient Method for Determining the Number of Components in Parafac Models," *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17, 274–286. [8]

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. [5,12]

Chen, X., Yang, Y. (2021), "Hanson–Wright Inequality in Hilbert Spaces With Application to K-Means Clustering for Non-Euclidean Data," *Bernoulli*, 27, 586–614. [9]

Chiles, J.-P., and Delfiner, P. (2009), *Geostatistics: Modeling Spatial Uncertainty* (Vol. 497), Hoboken, NJ: Wiley. [10]

Chollet, F. (2015), "Keras," Available at *https://github.com/fchollet/keras*. [12]

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997), "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science*, 278, 680–686. [2]

D'Souza-Schorey, C., and Clancy, J. W. (2012), "Tumor-Derived Microvesicles: Shedding Light on Novel Microenvironment Modulators and Prospective Cancer Biomarkers," *Genes & Development*, 26, 1287–1299. [1]

Dutilleul, P. (1999), "The MLE Algorithm for the Matrix Normal Distribution," *Journal of Statistical Computation and Simulation*, 64, 105–123. [1,2]

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, pp. 14863–14868. [2]

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [2,11]

Fu, Y., Matsushima, S., and Yamanishi, K. (2019), "Model Selection for Non-Negative Tensor Factorization With Minimum Description Length. *Entropy*, 21, 632. [8]

Gaetan, C., and Guyon, X. (2010), "Second-Order Spatial Models and Geostatistics," in *Spatial Statistics and Modeling*, New York, NY: Springer, pp. 1–52. [10]

Goutte, C., and Amini, M.-R. (2010), "Probabilistic Tensor Factorization and Model Selection," *Tensors, Kernels, and Machine Learning (TKLM 2010)*, pp. 1–4. [8]

Greenewald, K., Zhou, S., and Hero, A. (2019), "Tensor Graphical Lasso (TeraLasso)," *Journal of the Royal Statistical Society* , Series B, 81, 901–931. [4]

Hinrichs, C., Singh, V., Xu, G., Johnson, S. C., and Initiative, A. D. N. (2011), "Predictive Markers for Ad in a Multi-Modality Framework: An Analysis of MCI Progression in the ADNI Population," *Neuroimage*, 55, 574–589. [1]

Hitchcock, F. L. (1927), "The Expression of a Tensor or a Polyadic as a Sum of Products," *Journal of Mathematics and Physics*, 6, 164–189. [3,5]

Hoff, P. D. (2011), "Separable Covariance Arrays Via the Tucker Product, With Applications to Multivariate Relational Data," *Bayesian Analysis*, 6, 179–196. [2]

Kolda, T. G. (2006), "Multilinear Operators for Higher-Order Decompositions," Technical report, Albuquerque, NM/Livermore, CA: Sandia National Laboratories. [3]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [3]

Kolda, T. G., and Plantenga, T. (2014), "Tensor Rank Prediction Via Cross–Validation," Technical report, Livermore, CA: Sandia National Lab.(SNL-CA). [8]

Lampert, C. H., Ralaivola, L., and Zimin, A. (2018), "Dependency-Dependent Bounds for Sums of Dependent Random Variables," arXiv: 1811.01404. [10]

Li, X., Xu, D., Zhou, H., and Li, L. (2018a), "Tucker Tensor Regression and Neuroimaging Analysis," *Statistics in Biosciences*, 10, 520–545. [2]

Li, X., Zhou, Y., Zhu, Z., Liang, L., Yu, B., and Cao, W. (2018b), "Mapping Annual Urban Dynamics (1985–2015) Using Time Series of Landsat Data," *Remote Sensing of Environment*, 216, 674–683. [15]

Lindquist, M. A. (2008), "The Statistical Analysis of fMRI Data," *Statistical Science*, 23, 439–464. [1]

Liu, J., and Calhoun, V. D. (2014), "A Review of Multivariate Analyses in Imaging Genetics," *Frontiers in Neuroinformatics*, 8, 29. [1]

Madrid-Padilla, O. H. and Scott, J. (2017), "Tensor Decomposition With Generalized Lasso Penalties," *Journal of Computational and Graphical Statistics*, 26, 537–546. [4]

Manceur, A. M., and Dutilleul, P. (2013), "Maximum Likelihood Estimation for the Tensor Normal Distribution: Algorithm, Minimum Sample Size, and Empirical Bias and Dispersion," *Journal of Computational and Applied Mathematics*, 239, 37–49. [2]

Matérn, B. (2013), *Spatial Variation* (Vol. 36), New York: Springer Science & Business Media. [12]

MATLAB Image Process Toolbox (2018a), *Matlab Image Process Toolbox* (Ver10.2), Natick, MA: The MathWorks. [13]

Meinshausen, N., and Yu, B. (2009), "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," *The Annals of Statistics*, 37, 246–270. [10]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of M-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [10]

O'Brien, J., O'Keefe, K., LaViolette, P., DeLuca, A., Blacker, D., Dickerson, B., and Sperling, R. (2010), "Longitudinal fMRI in Elderly Reveals Loss of Hippocampal Activation With Clinical Decline," *Neurology*, 74, 1969–1976. [15]

Raskutti, G., Yuan, M., and Chen, H. (2019), "Convex Regularization for High-Dimensional Multiresponse Tensor Regression," *The Annals of Statistics*, 47, 1554–1584. [10]

Schlather, M. (1999), "An Introduction to Positive Definite Functions and to Unconditional Simulation of Random Fields," Technical report at 99-10, Lancaster: Dept. of Mathematics and Statistics, Lancaster University. [10]

Shah, P., Rao, N., and Tang, G. (2015), "Sparse and Low-Rank Tensor Decomposition," in *Advances in Neural Information Processing Systems*, 28, 2548–2556. [4]

Soltanian-Zadeh, S., Sahingur, K., Blau, S., Gong, Y., and Farsiu, S. (2019), "Fast and Robust Active Neuron Segmentation in Two-Photon Calcium Imaging Using Spatiotemporal Deep Learning," *Proceedings of the National Academy of Sciences*, 116, 8554–8563. [15]

Sun, W. W., and Li, L. (2019), "Dynamic Tensor Clustering," *Journal of the American Statistical Association*, 114, 1894–1907. [2,4,8]

Sun, W. W., Lu, J., Liu, H., and Cheng, G. (2017), "Provable Sparse Tensor Decomposition," *Journal of the Royal Statistical Society*, Series B, 3, 899–916. [4,7]

Tang, X., Bi, X., and Qu, A. (2019), "Individualized Multilayer Tensor Learning with an Application in Imaging Analysis," *Journal of the American Statistical Association*, pages 1–26. [1,2,8,10,12,15]

Taylor, D. D., and Gercel-Taylor, C. (2008), "Microrna Signatures of Tumor-Derived Exosomes as Diagnostic Biomarkers of Ovarian Cancer," *Gynecologic Oncology*, 110, 13–21. [1]

Tian, T. S. (2010), "Functional Data Analysis in Brain Imaging Studies," *Frontiers in Psychology*, 1, 35. [1]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness Via the Fused Lasso," *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [4]

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. [1,13,14]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31:279–311. [3]

Wackernagel, H. (2003), "Examples of Covariance Functions," in *Multivariate Geostatistics: An Introduction with Applications*, Berlin: Springer Berlin Heidelberg, pp. 57–61. [10]

Wang, L., Albera, L., Kachenoura, A., Shu, H., and Senhadji, L. (2014), "Canonical Polyadic Decomposition of Third-Order Semi-Nonnegative Semi-Symmetric Tensors Using LU and QR Matrix Factorizations," *EURASIP Journal on Advances in Signal Processing*, 2014, 150. [7]

Wang, M., and Li, L. (2020), "Learning From Binary Multiway Data: Probabilistic Tensor Decomposition and Its Statistical Optimality," *Journal of Machine Learning Research*, 21, 1–38. [2,9]

Witten, D. M., Tibshirani, R., and Hastie, T. (2009), "A Penalized Matrix Decomposition, With Applications to Sparse Principal Components and Canonical Correlation Analysis," *Biostatistics*, 10, 515–534. [7,8]

Wu, Y., Tan, H., Li, Y., Zhang, J., and Chen, X. (2019), "A Fused CP Factorization Method for Incomplete Tensors," *IEEE Transactions on Neural Networks and Learning Systems*, 30, 751–764. [4]

Xia, D., and Zhou, F. (2019), "The Sup-Norm Perturbation of HOSVD and Low Rank Tensor Denoising," *Journal of Machine Learning Research*, 20, 61–1. [2]

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., and Initiative, A. D. N. (2012), "Multi-Source Feature Learning for Joint Analysis of Incomplete Multiple Heterogeneous Neuroimaging Data," *NeuroImage*, 61, 622–632. [1]

Zhang, A., and Han, R. (2019), "Optimal Sparse Singular Value Decomposition for High-Dimensional High-Order Data," *Journal of the American Statistical Association*, 114, 1708–1725. [2,10]

Zhang, A., and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [2]

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. [1]

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression With Applications in Neuroimaging Data Analysis," *Journal of American Statistics Association*, 108, 229–239. [5,12]