Electronic Journal of Statistics
Vol. 16 (2022) 451–497
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1954

A general framework for tensor screening through smoothing*

Keqian Min and Qing Mai

Department of Statistics, Florida State University, Tallahassee, Florida 32306, U.S.A.
e-mail: km17g@my.fsu.edu; qmai@fsu.edu

Abstract: Screening is an important technique for analyzing high-dimensional data. Most screening tools have been developed for vectors and are marginal in the sense that each variable is evaluated individually at a time. Many multi-dimensional arrays (tensors) are generated nowadays. In addition to being high-dimensional, these data further have the tensor structure that should be exploited for more efficient analysis. Variables adjacent to each other in a tensor tend to be important or unimportant at the same time. Such information is ignored by marginal screening methods. In this article, we propose a general framework for tensor screening called smoothed tensor screening (STS). STS combines the strength of current marginal screening methods with tensor structural information by aggregating the information of its adjacent variables when evaluating one variable. STS is widely applicable since the statistical utility used in screening can be chosen based on the underlying model or data type of the responses and predictors. Moreover, we establish the SURE screening property for STS under mild conditions. Numerical studies demonstrate that STS has better performance than marginal screening methods.

MSC2020 subject classifications: 62P10, 62F07.
Keywords and phrases: Screening, smooth structure, SURE screening property, tensor.

Received October 2020.

1. Introduction

A large number of tensor datasets have been appearing in modern scientific research, attracting much attention to the analysis of such datasets. For example, advanced neuroimaging technology often generates imaging data in the form of tensors. Electroencephalography monitors brain activity at several locations for a period of time, resulting in 2-way tensors. Magnetic resonance imaging produces 3-way tensors as scans of brains. Similarly, functional magnetic resonance imaging data are 4-way tensors with the 4th dimension being the time domain. Tensor data are also frequently seen in other fields, such as computational biology, personalized recommendation, and image recognition analysis.

In principle, we could vectorize the tensors and then apply existing vector methods on the dataset. However, it is a wide consensus that vectorization could

*This project was supported in part by the grant CCF-1908969 from the U.S. National Science Foundation.

lead to loss of efficiency in analysis, as the tensor structure contains valuable information that usually cannot be adequately modeled by existing vector models. In this article, we focus on an especially important tensor structure, the smoothness structure.

Oftentimes, tensor data are collected in a way such that elements close to each other are similar. Thus, the coefficients of them are smooth. For example, [38, 39, 53] argue that brain images often have the smoothness structure in that voxels adjacent to each other tend to be all important or unimportant at the same time, because biologically voxels in the same brain region usually function together. [59] consider tensor generalized linear models with low-rank and sparse coefficients, as they want to perform region selection instead of variable selection. [49] impose the smoothness structure to capture the dynamic nature of tensor. When analyzing a brain image data collected over time, they assume that the brain activities change smoothly along the time domain. More concretely, in Section 5.2 we consider an electroencephalography (EEG) data, in which voltage fluctuations are collected from 64 electrodes placed on the scalp, which are sampled at 256 Hz for 1 second. As a result, the EEG images are stored as 2-way tensors, with smoothness structure along rows (i.e., time points) and columns (i.e., locations of the electrodes). Vectorization of tensor data destroys such smoothness structure; the adjacent elements may no longer be close to each other after vectorization. Instead, it is highly desirable to preserve and leverage the tensor structure rather than to vectorize the data.

Another common property of tensor data is their high dimensionality. Many tensor data naturally have a large number of elements in them. For example, the EEG dataset in Section 5.2 is of dimension 256×64, with 16,384 elements in total. [24] and [42] analyzed the attention deficit hyperactivity disorder (ADHD) data of dimension $30 \times 36 \times 30$, with $32,400$ elements in total. Such high dimensionality calls for additional assumptions for accurate modeling. Therefore, many researchers borrow the popular sparsity assumption from high-dimensional statistics to enforce variable selection and thus reduce the model complexity in tensor data analysis [59, 48, 42, e.g].

However, these high-dimensional tensor methods generally take the penalized approach, in which we consider an optimization problem that combines an appropriate loss function and a sparsity-inducing penalty. When the dimension is excessively high, penalized methods can be time-consuming or computationally unstable [10]. In such cases, screening is widely regarded as a computationally efficient preprocessing tool. For vector data, a large family of screening methods have been proposed [10, 12, 18, 13, 8, 25, 35, 3, 19, 11, 36, 4, e.g]. These methods marginally rank the variables by some properly chosen screening utilities. Only the variables that appear marginally important are preserved for further analysis.

Although screening methods effectively mitigate the impact of high dimensionality on vector data, when applied to tensor data, they are typically incapable of utilizing the tensor data structure. Since the screening utilities are calculated on individual variables, screening does not recognize the aforementioned common spatial structure in tensor data, nor does it encourage the selection of

important regions. Ignoring this important piece of information could drastically decrease the efficiency of our statistical analysis, especially in the presence of the high dimensions and the relatively limited sample size in tensor data.

To tackle this challenge, we propose a general framework for tensor screening that explicitly takes advantage of the tensor structure. We incorporate the common smoothness structure in tensor data into the screening procedure, resulting in a smoothed tensor screening (STS) framework. When we calculate the screening utility for one variable, we not only consider this individual variable, but also aggregate the information from adjacent variables. Consequently, adjacent variables tend to receive similar rankings, and we encourage the selection of important regions instead of scattered elements. STS can be combined with any existing marginal screening method to exploit the tensor structure for better variable selection. Moreover, STS can be completed with the same order of computationally costs as the corresponding marginal screening method and thus preserves the most attractive feature of marginal screening. We show that STS enjoys the so-called SURE screening property that it preserves all the important variables with a probability tending to 1 under mild conditions.

After STS, refined analysis can be performed on the reduced set. One could use either vector-based methods or tensor-based methods for this purpose. There are many methods developed for tensor model fitting. For example, for regression problems, see [59, 20, 43, 53, 24, 57, 31]. For classification problems, see [32, 42]. Many of these methods utilize the low-rank assumption, which is related to tensor decomposition [5, 50, 28, 56, e.g]. Most model-fitting methods can be easily combined with STS, either directly or with slight data augmentation. Some of them consist of variable selection and thus can further exclude more variables from the dataset. Others may not perform variable selection on their own. STS is a convenient way to add variable selection to these methods, besides boosting their computation efficiency. Our numerical studies demonstrate superior performance for STS combined with many popular model-fitting methods.

The rest of this article is organized as follows. In Section 2, we start with a review of marginal screening. We also introduce some useful tensor notation. Section 3 presents the procedure of STS and the analysis afterwards. The SURE screening property is established in Section 4. In Section 5, we present simulation results as well as a real data analysis example. Section 6 summarizes our contributions and discusses some future research directions. Additional numerical studies and technical proofs are given in the Appendix.

2. Background

2.1. Marginal screening

We first briefly review marginal screening for vector data, as the main purpose of this article is to generalize these methods to tensor data. Consider a random pair $\{\mathbf{U}, Y\}$, where $\mathbf{U} \in \mathbb{R}^p$ is a p-dimensional vector of predictors, and Y is

the univariate response that can be either continuous or discrete. Some works consider multivariate Y, but we focus on the univariate response case for ease of presentation. Nevertheless, our framework easily extends to the multivariate response. We observe n independent and identically distributed copies of the random pair, denoted as $\left(\mathbf{U}^i, Y^i\right), i = 1, \ldots, n$. Consider high-dimensional problems where p is much larger than n. In this scenario, fitting models with all variables often leads to a drastic loss of efficiency and accuracy. Rather, statisticians often perform variable selection under the celebrated sparsity assumption. The sparsity assumption means that only a small subset of variables are related to the response Y. More rigorously, we define the set of important variables [60] as

$$\mathcal{D} = \{j : F(y \mid \mathbf{U}) \text{ functionally depends on } U_j \text{ for some } y\},$$

where U_j is the jth feature of $\mathbf{U}$ and $F(y \mid \mathbf{U}) = \text{pr}(Y < y \mid \mathbf{U})$ is the conditional distribution function of Y given $\mathbf{U}$. The sparsity assumption implies that $|\mathcal{D}| \ll p$. Hence, if we can identify $\mathcal{D}$, estimation and prediction can be performed within a much lower-dimensional space.

Marginal screening is a family of computationally efficient techniques to identify $\mathcal{D}$. They are usually carried out in the following procedure. First, we choose a proper screening utility ϕ_{nj} that measures the marginal dependence between U_j and Y such that a larger value of ϕ_{nj} indicates stronger dependence. For example, if Y is continuous, we can choose ϕ_{nj} to be the absolute value of the Pearson correlation between Y and U_j [10]. If Y is binary, ϕ_{nj} could be the absolute value of the t-statistic of U_j across the two levels of Y [7]. As mentioned in Section 1, more sophisticated utilities have been proposed in the literature as well to accommodate more complicated statistical models. With a chosen screening utility, all the variables are ranked by their corresponding ϕ_{nj}. Only the variables with the highest ranks are selected, i.e., we select the set

$$\widehat{\mathcal{S}}_\phi\left(d_n\right) = \{j : \phi_{nj} \text{ is amongst the first } d_n \text{ largest of all}\},$$

where d_n is a positive integer predefined by users. Since most penalized methods can only deal with $o(n)$ variables, the common choices for d_n are n and $\lceil n/\log n \rceil$, where $\lceil a \rceil = \min\{i : i \geq a \text{ and } i \text{ is an integer}\}$ for $a > 0$.

Since we only fit the model on $\widehat{\mathcal{S}}_\phi(d_n)$, it is of utmost importance that $\widehat{\mathcal{S}}_\phi(d_n)$ contains all the important variables in $\mathcal{D}$. More formally, a screening method is said to enjoy the SURE screening property if $\mathcal{D} \subseteq \widehat{\mathcal{S}}_\phi\left(d_n\right)$. Most existing screening methods enjoy the SURE screening property under two types of interpretable conditions. Define ϕ_j as the population counterpart of ϕ_{nj}. The two conditions are:

Condition (V1). *There exists $\mathcal{S}$ such that $\mathcal{D} \subseteq \mathcal{S}$ and $\delta_{\mathcal{S}} = \min_{j \in \mathcal{S}}\{\phi_j\} - \max_{j \in \mathcal{S}^c}\{\phi_j\} > 0$.*

Condition (V2). *There exist a constant $\epsilon_0 > 0$ and a monotonically decreasing function ζ_n such that for any $0 < \epsilon < \epsilon_0$,*

$$\text{pr}\left(\max_{1 \leq j \leq p} |\phi_{nj} - \phi_j| > \epsilon\right) \leq p\zeta_n(\epsilon).$$

Condition (V1) guarantees the validity of screening on the population level; if we knew the true model, ϕ_j should provide a reasonable ranking such that the important variables are ranked higher than the unimportant ones. Condition (V2) requires ϕ_{nj} to be accurate approximations of ϕ_j, such that we can preserve the ranking on the sample level. Condition (V2) is replaced with suitable lower-level conditions for specific screening methods (see, e.g., [13, 25, 35]).

2.2. Tensor notation

We introduce some tensor notation that will be used throughout the rest of the article. See [21] for a review of notation and operations of tensor. A tensor is a multi-dimensional array, and its dimension is called the order or ways of the tensor. An R-dimensional array $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ is a tensor of order R. Vectors and matrices can also be viewed as order-one ($R = 1$) tensors and order-two ($R = 2$) tensors. The order of a tensor is also known as modes. The mode-r product of tensor $\mathbf{A}$ with a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{d \times p_r}$ is defined as $\mathbf{A} \times_r \boldsymbol{\alpha}$ and it yields a tensor of size $p_1 \times \cdots \times p_{r-1} \times d \times p_{r+1} \times \cdots \times p_R$. The Tucker decomposition of $\mathbf{A}$, defined as $\mathbf{A} = \mathbf{C} \times_1 \mathbf{G}_1 \cdots \times_R \mathbf{G}_R$, can decompose $\mathbf{A}$ into the product of a core tensor $\mathbf{C} \in \mathbb{R}^{d_1 \times \cdots \times d_R}$ and R factor matrices $\mathbf{G}_r \in \mathbb{R}^{p_r \times d_r}$, $r = 1, \ldots, R$. The Tucker decomposition is often written in a shorthand notation, $\mathbf{A} = [\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_R]\!]$. If all elements in $\mathbf{A}$ independently follow the standard normal distribution and $\mathbf{X} = \boldsymbol{\mu} + [\![\mathbf{A}; \boldsymbol{\Sigma}_1^{1/2}, \ldots, \boldsymbol{\Sigma}_R^{1/2}]\!]$, then $\mathbf{X}$ follows a tensor normal (TN) distribution, denoted as $\mathbf{X} \sim TN(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R)$. If $R = 2$, the tensor normal distribution reduces to the matrix normal (MN) distribution [17].

3. Smoothed tensor screening

3.1. The proposed smoothed tensor screening procedure

We consider screening on tensor data. We are interested in the random pair $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ is a R-dimensional tensor predictor and Y is the univariate response that can be either continuous or discrete. The tensor predictor $\mathbf{X}$ has $p = \prod_{r=1}^{R} p_r$ elements, which is often an intimidating number. We observe a random sample $\{\mathbf{X}^i, Y^i\}, i = 1, \ldots, n$, where n is much smaller than p.

We want to perform screening on the observed data to reduce the number of variables. Also, recall that we want to leverage the tensor structure for better screening results. To this end, we assume that variables adjacent to each other tend to be important or unimportant at the same time. Let $\mathcal{J} = (j_1, \ldots, j_R)$ be an index and $X_{\mathcal{J}}$ be the $\mathcal{J}$th element in $\mathbf{X}$. The tensor version definition of the set of important variables is

$$\mathcal{D} = \{\mathcal{J} : F(y \mid \mathbf{X}) \text{ functionally depends on } X_{\mathcal{J}} \text{ for some } y\}.$$

For high-dimensional tensors, we propose a general smoothed tensor screening (STS) framework for $(\mathbf{X}, Y)$. STS consists of the following three steps.

First, we choose an appropriate measurement of the dependence between Y and $X_{\mathcal{J}}$, $\phi_{n\mathcal{J}}$, for each $\mathcal{J}$. Most existing marginal utilities can be candidates as long as their corresponding models are reasonable. For example, if we believe that Y and $\mathbf{X}$ are related through a linear regression model, $\phi_{n\mathcal{J}}$ could be the Pearson correlation [10]. If a generalized linear model is suitable, we can take $\phi_{n\mathcal{J}}$ to be the coefficient of the marginal generalized linear model [13]. If we wish to perform screening in a model-free fashion, distance correlation can be applied [25]. We will demonstrate our proposed framework with three popular choices in later sections.

Second, we exploit the tensor structure by incorporating the neighborhood information. For each $\mathcal{J} = (j_1, \ldots, j_R)$, define $\Omega_{\mathcal{J}} = \{\mathcal{I} = (i_1, \ldots, i_R) : |i_r - j_r| \leq 1, r = 1, \ldots, R\} \backslash \{\mathcal{J}\}$. Apparently, $\Omega_{\mathcal{J}}$ contains all the predictors adjacent to $X_{\mathcal{J}}$. Then we obtain the smoothed screening utility

$$\phi_{n\mathcal{J}}^{\text{Smooth}} = \phi_{n\mathcal{J}} + c \cdot \overline{\phi}_{n\Omega_{\mathcal{J}}}, \tag{3.1}$$

where $\overline{\phi}_{n\Omega_{\mathcal{J}}} = \frac{1}{|\Omega_{\mathcal{J}}|} \sum_{\mathcal{I} \in \Omega_{\mathcal{J}}} \phi_{n\mathcal{I}}$ is the average dependence across the neighborhood of $X_{\mathcal{J}}$ and $c \geq 0$ is the user-specified weight. Since $\phi_{n\mathcal{J}}^{\text{Smooth}}$ is the weighted sum of $\phi_{n\mathcal{J}}$ and $\overline{\phi}_{n\Omega_{\mathcal{J}}}$, it combines the information from $X_{\mathcal{J}}$ with that from its neighbors. The weight c determines the level of smoothness in $\phi_{n\mathcal{J}}^{\text{Smooth}}$. When $c = 0$, $\phi_{n\mathcal{J}}^{\text{Smooth}}$ reduces to the marginal screening utility $\phi_{n\mathcal{J}}$ that does not promote smoothness. As we increase c, $\phi_{n\mathcal{J}}^{\text{Smooth}}$ becomes smoother across $\mathcal{J}$. For variable $X_{\mathcal{J}}$, if $c = |\Omega_{\mathcal{J}}|$, all adjacent variables are treated equally with $X_{\mathcal{J}}$ itself.

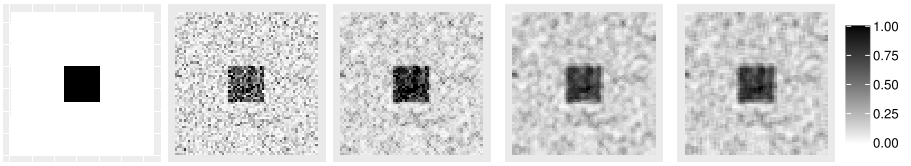
If desired, c can be chosen by cross-validation. However, we discover that this is generally not necessary if smoothness structure exists. The screening results are not sensitive to the choice of c, as long as it is in a reasonable range. Define $\omega = \max_{\mathcal{J}} |\Omega_{\mathcal{J}}|$. Generally, if $p_r \geq 3$ for $r = 1, \ldots, R$, we have $\omega = 3^R - 1$. If $\omega/2 \leq c \leq \omega$, the screening results are roughly constant in all our simulation models that satisfy the smoothness assumption.

Finally, we rank the variables by $\phi_{n\mathcal{J}}^{\text{Smooth}}$ and select the following subset

$$\widehat{\mathcal{S}}_{\phi}^{\text{Smooth}}(d_n) = \left\{\mathcal{J} : \phi_{n\mathcal{J}}^{\text{Smooth}} \text{ is amongst the first } d_n \text{ largest of all}\right\}. \tag{3.2}$$

Following the convention in marginal screening, we could set d_n to be $\lceil n/\log n \rceil$ or n.

Our proposed screening procedure is different from marginal screening in that it smoothes the screening utilities across the locations over the tensor. Hence, we refer to it as smoothed tensor screening (STS). STS is a general framework for tensor screening because it can be combined with any marginal screening method to achieve smoothed variable selection on tensor data. By summing $\phi_{n\mathcal{J}}$ with its neighbors, we utilize the natural spatial structure in tensor data to obtain better variable selection. Moreover, STS has the same order of computation cost as marginal screening. For example, if the computation cost for $\phi_{n\mathcal{J}}$ is $O(n^{\gamma})$ for some $\gamma > 0$, then the computation cost for the corresponding marginal screening is $O(n^{\gamma} \cdot p)$. STS has an additional smoothing step, with the computation cost



(a) True signal (b) $c = 0$ (c) $c = 1$ (d) $c = 4$ (e) $c = 8$

FIG 1. *Heatmaps of $T_{n\mathcal{J}}^{Smooth}$ at different c. For each c, $T_{n\mathcal{J}}^{Smooth}$ is scaled to the range of $[0, 1]$. When $c = 0$, the procedure is equivalent to marginal screening that ignores the tensor structure and does not effectively separate the signals and noises. After smoothing the statistics, we have a better recovery of the true signal block.*

of $O(\omega p) = o(n^{\gamma} \cdot p)$ as n becomes large. Hence, the total computation cost for STS remains the same as the corresponding marginal screening in use.

We further present a toy example to illustrate the procedure of STS.

Example 1. Consider a binary classification problem where the response $Y = 1$ or 2 with equal probabilities and the predictor $\mathbf{X} \in \mathbb{R}^{64\times 64}$. We generate $\mathbf{X}$ from the matrix normal distribution $\mathbf{X} \mid (Y = k) \sim MN(\boldsymbol{\mu}_k, \mathbf{I}_{64}, \mathbf{I}_{64})$. Define $\mathcal{D} = \{(i, j) : 25 \le i, j \le 40\}$. In the first class, $\boldsymbol{\mu}_{1,\mathcal{D}} = 0.35$ and $\boldsymbol{\mu}_{1,\mathcal{D}^c} = 0$. In the second class, $\boldsymbol{\mu}_2 = 0$. For each class, we simulate 150 samples, which is a small sample size compared to the 4,096 elements in $\mathbf{X}$.

Our model is known as a tensor discriminant analysis model [42]. It is similar to a discriminant analysis model, in which the t-statistics can be used for marginal screening [7]. Denote $T_{n\mathcal{J}}$ as the absolute value of the t-statistic calculated on $(X_{\mathcal{J}}, Y)$. Then we compute $T_{n\mathcal{J}}^{\text{Smooth}}$ as in (3.1) with $c = 0, 1, 4, 8$, where we note that $c = 0$ corresponds to marginal screening and 8 is the maximum number of neighbors in our model. See Figure 1 for the heatmaps of the resulting $T_{n\mathcal{J}}^{\text{Smooth}}$. Since the signals in our model are relatively weak, marginal screening (Panel (b), $c = 0$) cannot effectively identify the important variables. Many important elements have small $T_{n\mathcal{J}}$, while many unimportant elements have misleadingly large $T_{n\mathcal{J}}$. However, as we exploit the tensor structure by increasing c, the signal block stands out clearly. While the block for $c = 1$ (Panel (c)) is somewhat blurry, it becomes much more distinctive for $c = 4, 8$ (Panels (d) & (e)). Hence, by encouraging smoothness, STS is more efficient in distinguishing the important elements from the unimportant ones. Moreover, Panels (d) & (e) look similar, which demonstrates our earlier claim that STS is not sensitive to the choice of c as long as it is reasonably large.

With the screening utilities, STS chooses the top d_n predictors. We let $d_n = n$ and plot the proportion of selected active predictors and that of selected inactive predictors, corresponding to $c = 0, 1, \ldots, 8$ in Figure 2. With the increase of c, most active predictors are selected and very few inactive predictors are selected. Also, the proportions are very stable for $c \ge 2$, indicating that there is no need to finely tune c under the smoothness assumption.

Finally, we comment on the robustness of STS. STS is proposed to exploit the

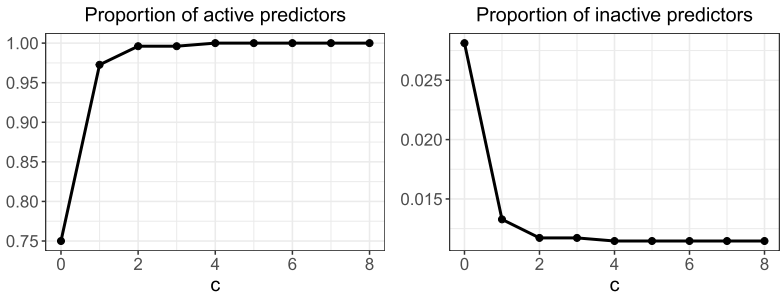


FIG 2. *Proportions of active and inactive predictors that are picked out at different weights c. After smoothing the statistics, more active predictors and fewer inactive predictors are kept.*

smoothness structure in the data. If there is no smoothness, STS is not recommended, as the screening results are expected to be not as accurate as marginal screening. However, STS is resistant to partial violations of the smoothness assumption. By our construction, STS is most efficient when the tensor is smooth along all modes. But sometimes in practice, the tensor is only smooth along some modes. For example, for a matrix predictor, it could be the case that the tensor is only smooth along the rows, but not the columns. Our numerical studies show that STS still works well in this partially smooth scenario. Moreover, recall that, in smooth models, we recommend $\omega/2 \leq c \leq \omega$. Among these choices, $c = \omega/2$ is the most robust to possible violations of the smoothness assumption. See Section 5.1 for empirical evidence of our discussion.

3.2. Other possible approaches for smoothed tensor screening

Compared to marginal screening, the most important innovation for STS is to aggregate information from neighbors to take advantage of the smoothness structure. We achieve this goal by taking the weighted average of an element and its immediately adjacent neighbors. However, there are other possible ways to smooth the screening utilities. We discuss two possibilities.

First, STS only uses the immediately adjacent elements, resulting in neighborhoods of size 3 along each mode. One could expand the neighborhoods to obtain more smoothed results. We investigate the effect of neighborhood size in Appendix C. We find that if there is strong smoothness structure, a neighborhood of size 5 or 7 (i.e., two or three adjacent elements are used instead of one along each direction) could have slightly better results than 3, but the improvement is minimal. Further increasing the neighborhood size no longer helps, and overly large neighborhood will eventually hurt the performance of STS. Moreover, even for the size of 5, there is a notable drop in the robustness. In other words, when the smoothness assumption is not fully satisfied, STS has much worse results with a neighborhood size of 5 compared to 3. Therefore, larger neighborhoods should only be used when the true model is very smooth. Otherwise, the neighborhood size of 3 is safer.

Second, there are other methods for the smoothing step in STS. For example, Gaussian filter is a very popular imaging processing technique [16]. Although Gaussian filter is often employed to smooth the data itself instead of screening utilities, it is straightforward to be applied in the screening context. With Gaussian filter, we still take the weighted average of the screening utilities within a neighborhood, but the weights are generated by a Gaussian kernel instead of being a constant as in STS. See Appendix D for an introduction of Gaussian filter and the comparison between it and STS. In STS, the constant c controls the degree of smoothness. In Gaussian filter, the standard deviation of the Gaussian kernel has the same purpose. When the standard deviation is chosen well, Gaussian filter performs almost identically to STS. Hence, throughout the rest of this article, we restrict our attention to STS to avoid redundancy.

3.3. Analysis after smoothed tensor screening

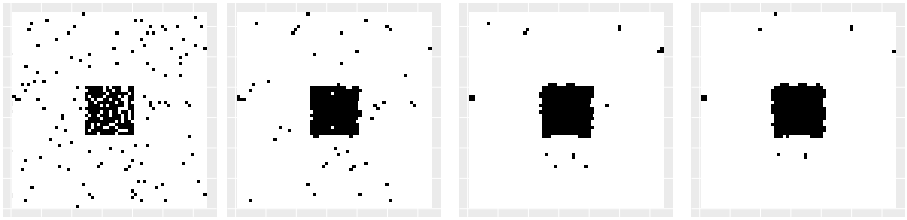
After STS, we perform model fitting on the selected variables. This is usually straightforward to do if we choose a vector-based method. We simply apply a vector-based method on $(Y, \mathbf{X}_{\mathrm{STS}})$, where $\mathbf{X}_{\mathrm{STS}}$ contains all the elements preserved by STS. However, it might be difficult to directly use a tensor-based analysis tool, because the reduced set $\mathbf{X}_{\mathrm{STS}}$ may no longer be a tensor. Hence, if we hope to use a tensor-based analysis tool, we first augment the reduced set into a tensor. Given $\widehat{\mathcal{S}}_{\phi}^{\mathrm{Smooth}}(d_n)$, we find the index set for each mode. Specifically, for $r = 1, \ldots, R$, define

$$M_r = \{k_r : \mathcal{J} = (j_1, \ldots, k_r, \ldots, j_R) \in \widehat{\mathcal{S}}_{\phi}^{\mathrm{Smooth}}(d_n) \text{ for some } (j_1, \ldots, j_{r-1}, j_{r+1}, \ldots, j_R)\}. \tag{3.3}$$

Then we keep all the variables in $M = \{\mathcal{J} = (j_1, \ldots, j_R) : j_1 \in M_1, \ldots, j_R \in M_R\}$ to form the new predictor $\widetilde{\mathbf{X}} \in \mathbb{R}^{|M_1| \times \ldots \times |M_R|}$. The resulting $\widetilde{\mathbf{X}}$ is the smallest tensor that contains $\mathbf{X}_{\mathrm{STS}}$. Further analysis can be performed on $\widetilde{\mathbf{X}}$. We continue to use the model setting in Example 1 to illustrate this augmentation step.

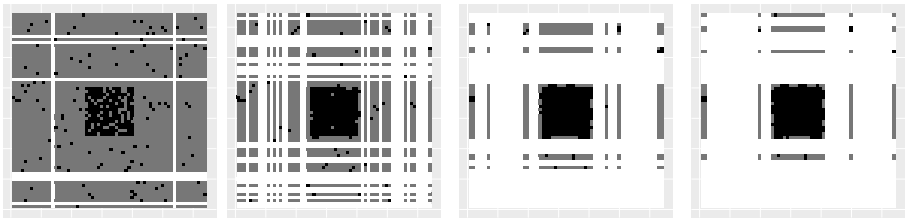
Example 1. (Cont'd) Under the same setting in Example 1, we plot the selected variables in Figure 3. Apparently, the selected variables (black) no longer form a tensor. We augment the selection results by further including the grey elements in Figure 4. The black and grey elements together form a smaller tensor. When c is reasonably large and STS works well, the augmentation does not include too many extra variables.

When the signal in tensor is smooth and STS produces an accurate selection result, the augmented data $\widetilde{\mathbf{X}}$ is often a small tensor, as is the case in Example 1. But if one is concerned with the inclusion of extra variables in the augmentation, the screening utilities can be combined with the penalty to filter out these variables. For example, denote $\mathbf{B} \in \mathbb{R}^{|M_1| \times \cdots \times |M_R|}$ as the parameter



(a) $c = 0$ (b) $c = 1$ (c) $c = 4$ (d) $c = 8$

FIG 3. *Selected variables after STS/marginal screening. Each black dot denotes a selected variable, and the white regions contain the variables identified to be unimportant.*



(a) $c = 0$ (b) $c = 1$ (c) $c = 4$ (d) $c = 8$

FIG 4. *Selected variables after augmentation. Black dots are variables selected by STS and grey dots are variables included to maintain the tensor form of the data. White dots are not selected.*

of interest and $\sum_{\mathcal{J}} \lambda_{\mathcal{J}} P(B_{\mathcal{J}})$ as the penalty function we choose. Then we can let $\lambda_{\mathcal{J}} = \lambda / \phi_{n\mathcal{J}}^{\text{Smooth}}$ for some $\lambda > 0$ such that the added variables are more heavily penalized than the ones preserved by screening. However, we do not observe significant improvement in this penalized procedure because $\widetilde{\mathbf{X}}$ usually only includes a small number of extra variables when STS is performed with a suitable c.

3.4. Smoothed tensor screening with three popular screening methods

In what follows, we present three examples for STS combined with three popular marginal screening utilities: the t-statistic, the maximum marginal likelihood estimator (MMLE), and the distance correlation. The t-statistic is a natural screening statistic for classification problems, MMLE works for the generalized linear model, and distance correlation is a successful model-free screening method. All these methods were originally proposed for vector data, but their direct generalizations to tensor data are straightforward. We first give a short review of the three statistical utilities in their original model setting, and then we extend them to tensor data. Throughout the rest of this section, we use

$\mathbf{U} \in \mathbb{R}^p$ to denote the vector predictors and $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ to denote the tensor predictors. Our observations are denoted as $\{\mathbf{X}^i, Y^i\}_{i=1}^n$.

3.4.1. The t-statistic

Consider a binary classification problem where $Y \in \{1, 2\}$ denotes the class label and $\mathbf{U} \in \mathbb{R}^p$ denotes the vector predictor. When $Y = k, k = 1, 2$, we assume that

$$\mathbf{U}_k = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the mean vector of class k and $\boldsymbol{\epsilon}_k \in \mathbb{R}^p$ is the error term, in which each element has mean zero. [7] proposed to calculate the t-statistic of each U_j across the two levels of Y and rank the importance of U_j by the magnitude of its corresponding t-statistic.

For tensor data, consider the class label $Y \in \{1, 2\}$ and the R-way tensor predictor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_R}$. When $Y = k$, we assume that

$$\mathbf{X}_k = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_k, \tag{3.4}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ is the mean tensor of class k and $\boldsymbol{\epsilon}_k \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ is the error term where each element has a mean of zero. We continue to use the two sample t-statistic on tensor data. In a dataset with n observations, assume that there are n_k samples within the class $Y = k$. We calculate

$$\phi_{n\mathcal{J}}^t = \frac{|\overline{X}_{1\mathcal{J}} - \overline{X}_{2\mathcal{J}}|}{n^{1/2}(H_{1\mathcal{J}}^2/n_1 + H_{2\mathcal{J}}^2/n_2)^{1/2}}, \tag{3.5}$$

where $\overline{X}_{k\mathcal{J}} = \frac{1}{n_k} \sum_{Y^i=k} X_{\mathcal{J}}^i$ and $H_{k\mathcal{J}}^2 = \frac{1}{n_k - 1} \sum_{Y^i=k} \left(X_{\mathcal{J}}^i - \overline{X}_{k\mathcal{J}}\right)^2$. Note that $\phi_{n\mathcal{J}}^t$ is the t-statistic divided by $n^{1/2}$. Since $\phi_{n\mathcal{J}}^t$ and $n^{1/2}\phi_{n\mathcal{J}}^t$ give us the same ranking of predictors, they are the same for the sake of screening. But such rescaling helps us define the population statistical utility

$$\phi_{\mathcal{J}}^t = \frac{|\mu_{1\mathcal{J}} - \mu_{2\mathcal{J}}|}{(\sigma_{1\mathcal{J}}^2/\pi_1 + \sigma_{2\mathcal{J}}^2/\pi_2)^{1/2}}, \tag{3.6}$$

where $\mu_{k\mathcal{J}}$ and $\sigma_{k\mathcal{J}}^2$ are the true population mean and variance of the $\mathcal{J}$th variable in class k and $\pi_k = \text{pr}(Y = k)$.

With $\phi_{n\mathcal{J}}^t$, we apply STS by calculating

$$\phi_{n\mathcal{J}}^{t.\text{STS}} = \phi_{n\mathcal{J}}^t + \frac{c}{|\Omega_{\mathcal{J}}|} \sum_{\mathcal{I} \in \Omega_{\mathcal{J}}} \phi_{n\mathcal{I}}^t.$$

The variables are then ranked by $\phi_{n\mathcal{J}}^{t.\text{STS}}$ and we select the following subset

$$\widehat{\mathcal{S}}_{\phi}^{t.\text{STS}}(d_n) = \left\{\mathcal{J} : \phi_{n\mathcal{J}}^{t.\text{STS}} \text{ is amongst the first } d_n \text{ largest of all}\right\}.$$

We refer to this procedure as STS-t screening. Similar to the t screening, STS-t screening is suitable when the response Y is a categorical variable. However,

STS-t screening is designed for tensor predictors instead of vector predictors. By taking into account the smoothness structure of tensor data, we achieve more efficient screening. After STS-t screening, we could apply a wide range of model-fitting methods. For example, if we focus on $\mathbf{X}_{\widehat{\mathcal{D}}}$, we can apply sparse discriminant analysis methods for vector data, such as [2, 6, 45, 54, 9, 37, 55, 34]. Alternatively, with the data augmentation in Section 3.3, we can apply tensor classification methods such as covariate-adjusted tensor classification (CATCH, [42]) and tensor regression based on CP decomposition (CP-GLM, [59]). Note that the tensor methods cannot be easily combined with marginal t-screening because t-screening does not honor the tensor structure.

3.4.2. The marginal maximum likelihood estimator

Suppose that the response Y is from an exponential family whose probability density function has the canonical form

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\},$$

where $b(\cdot), c(\cdot)$ are known functions and θ is an unknown function. Suppose that the predictor $\mathbf{U}$ is a p-dimensional vector and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is the parameter. The following generalized linear model is assumed:

$$\mathrm{E}(Y \mid \mathbf{U} = \mathbf{u}) = b'(\theta(\mathbf{u})) = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j u_j\right),$$

where $g = (b')^{-1}$ is the link function. Under this model, [13] proposed the screening method based on the maximum marginal likelihood estimator (MMLE), which is obtained from componentwise regression. The marginal estimator $\widehat{\beta}_j^M$ is defined by

$$\left(\widehat{\beta}_{j,0}^M, \widehat{\beta}_j^M\right) = \arg\min_{\beta_0,\beta_j} \left\{\frac{1}{n}\sum_{i=1}^{n} l\left(\beta_0 + \beta_j U_j^i, Y^i\right)\right\},$$

where $l(Y;\theta) = -[\theta Y - b(\theta) - \log c(Y)]$. Then the marginal screening utility is chosen to be $\phi_{nj}^{\mathrm{GLM}} = |\widehat{\beta}_j^M|$.

For tensor predictor $\mathbf{X}$, we assume a similar model:

$$\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}) = b'(\theta(\mathbf{x})) = g^{-1}\left(\beta_0 + \sum_{\mathcal{J}} \beta_{\mathcal{J}} x_{\mathcal{J}}\right).$$

Define the marginal estimator

$$\left(\widehat{\beta}_{\mathcal{J},0}^M, \widehat{\beta}_{\mathcal{J}}^M\right) = \arg\min_{\beta_0,\beta_{\mathcal{J}}} \left\{\frac{1}{n}\sum_{i=1}^{n} l\left(\beta_0 + \beta_{\mathcal{J}} X_{\mathcal{J}}^i, Y^i\right)\right\}.$$

The marginal screening utility is chosen to be $\phi_{n\mathcal{J}}^{\mathrm{GLM}} = |\widehat{\beta}_{\mathcal{J}}^{M}|$. Its population version $\phi_{\mathcal{J}}^{\mathrm{GLM}}$ is similarly obtained by minimizing $\mathrm{E}(l(\beta_0 + \beta_{\mathcal{J}} X_{\mathcal{J}}, Y))$. To apply STS, we further calculate

$$\phi_{n\mathcal{J}}^{\mathrm{GLM.STS}} = \phi_{n\mathcal{J}}^{\mathrm{GLM}} + \frac{c}{|\Omega_{\mathcal{J}}|} \sum_{\mathcal{I} \in \Omega_{\mathcal{J}}} \phi_{n\mathcal{I}}^{\mathrm{GLM}}.$$

The variables are then ranked by $\phi_{n\mathcal{J}}^{\mathrm{GLM.STS}}$ and we select the following subset

$$\widehat{\mathcal{S}}_{\phi}^{\mathrm{GLM.STS}}(d_n) = \left\{ \mathcal{J} : \phi_{n\mathcal{J}}^{\mathrm{GLM.STS}} \text{ is amongst the first } d_n \text{ largest of all} \right\}.$$

We refer to this procedure as STS-GLM screening. For various types of response such as Y is binomial or normal, whenever the generalized linear model is suitable, we can apply the STS-GLM screening. Through using the smoothness structure of tensor data, STS-GLM screening is more efficient than the vector-based GLM screening on tensor data. After obtaining the screened data, we can apply vector-based methods such as [14], or, with the data augmentation step, we can also apply tensor regression methods such as CP-GLM [59].

3.4.3. Distance correlation screening

Distance correlation screening [25] is a model-free screening method that works for any statistical model. It uses distance correlation [51] to measure the dependence between the response and the predictor. We first briefly review distance correlation.

Consider two random variables V, W. Distance correlation can be calculated for a pair of vectors, but we only consider univariate random variables $V, W \in \mathbb{R}$ here for ease of presentation. The squared distance covariance is defined as:

$$\mathrm{dcov}^2(V, W) = \frac{1}{\pi^2} \int_{\mathbb{R}^2} \frac{|\alpha_{V,W}(s,t) - \alpha_V(s)\alpha_W(t)|^2}{s^2 t^2} \,\mathrm{d}\, t \,\mathrm{d}\, s, \tag{3.7}$$

where $\alpha_{V,W}$ is the joint characteristic function of (V, W), α_V is the characteristic function of V and α_W is the characteristic function of W. The distance correlation $\mathrm{dcorr}(V, W)$ between V and W is defined as

$$\mathrm{dcorr}(V, W) = \frac{\mathrm{dcov}(V, W)}{(\mathrm{dcov}(V, V)\,\mathrm{dcov}(W, W))^{1/2}}. \tag{3.8}$$

The distance correlation is a measurement of dependence between V, W because $\mathrm{dcorr}(V, W) = 0$ if and only if V and W are independent. In practice, with n samples, the distance covariance is estimated by

$$\begin{aligned} \widehat{\mathrm{dcov}}^2(V, W) \quad = \quad & \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |V^i - V^j||W^i - W^j| \\ & + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |V^i - V^j| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |W^i - W^j| \end{aligned}$$

$$-\frac{2}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|V^i-V^l||W^j-W^l|,$$

and $\widehat{\text{dcorr}}(V,W)$ can be obtained accordingly. In distance correlation screening, the estimated distance correlation is used as the screening utility.

To apply distance correlation screening with tensor predictors, we consider the response Y and the tensor predictor $\mathbf{X}\in\mathbb{R}^{p_1\times\cdots\times p_R}$. We define the marginal screening utility

$$\phi_{n\mathcal{J}}^{\text{DC}}=\widehat{\text{dcorr}}^{\,2}\left(X_{\mathcal{J}},Y\right).$$

To apply STS, we further compute the smoothed screening utility

$$\phi_{n\mathcal{J}}^{\text{DC.STS}}=\phi_{n\mathcal{J}}^{\text{DC}}+\frac{c}{|\Omega_{\mathcal{J}}|}\sum_{\mathcal{I}\in\Omega_{\mathcal{J}}}\phi_{n\mathcal{I}}^{\text{DC}}.$$

Then we keep the variables with the top d_n values of $\phi_{n\mathcal{J}}^{\text{DC.STS}}$, that is

$$\widehat{\mathcal{S}}_{\phi}^{\text{DC.STS}}\left(d_n\right)=\left\{\mathcal{J}:\phi_{n\mathcal{J}}^{\text{DC.STS}}\text{ is amongst the first }d_n\text{ largest of all}\right\}.$$

We refer to this procedure as STS-DC screening. Since the distance correlation screening is a model-free screening method, the STS-DC screening is also model-free. Almost all model-fitting methods can be applied after STS-DC screening. Moreover, STS-DC is a screening method for tensor predictors and is expected to achieve better screening results when there is a smoothness structure in the tensor.

4. Theoretical properties

4.1. A generic theorem

In this section, we present the theoretical properties of STS. Since STS is a general framework that can be combined with any screening utility, we first consider its properties for a generic utility $\phi_{n\mathcal{J}}$. Then we present more concrete results for STS with the three popular screening methods mentioned in Section 3.4. All our proofs are included in Appendix A.

For a generic $\phi_{n\mathcal{J}}$ and its smoothed version $\phi_{n\mathcal{J}}^{\text{Smooth}}$, we respectively define their population counterparts as $\phi_{\mathcal{J}}$ and

$$\phi_{\mathcal{J}}^{\text{Smooth}}=\phi_{\mathcal{J}}+c\cdot\overline{\phi}_{\Omega_{\mathcal{J}}},\tag{4.1}$$

where $\overline{\phi}_{\Omega_{\mathcal{J}}}=\frac{1}{|\Omega_{\mathcal{J}}|}\sum_{\mathcal{I}\in\Omega_{\mathcal{J}}}\phi_{\mathcal{I}}$. We make the following assumptions on $\phi_{\mathcal{J}}^{\text{Smooth}}$.

Condition (T1). *There exists $\mathcal{S}_1$ such that $\mathcal{D}\subseteq\mathcal{S}_1$ and*

$$\delta_{\mathcal{S}_1}^{Smooth}=\min_{\mathcal{J}\in\mathcal{S}_1}\left\{\phi_{\mathcal{J}}^{Smooth}\right\}-\max_{\mathcal{J}\in\mathcal{S}_1^c}\left\{\phi_{\mathcal{J}}^{Smooth}\right\}>0.$$

Condition (T2). *There exist a constant* $\epsilon_0 > 0$ *and a monotonically decreasing function* ζ_n *such that for any* $0 < \epsilon < \epsilon_0$*, we have*

$$\text{pr}\left(\max_{\mathcal{J}} |\phi_{n\mathcal{J}} - \phi_{\mathcal{J}}| > \epsilon\right) \leq (\prod_{r=1}^{R} p_r)\zeta_n(\epsilon).$$

Apparently Conditions (T1) & (T2) are similar to Conditions (V1) & (V2) introduced in Section 2.1 for marginal screening; see more discussion after Theorem 4.1. Recall that ω is the maximum number of neighbors, and c is the weight for neighbors. We have the following theorem.

Theorem 4.1. *Suppose that Condition (T2) holds. We have the following conclusions.*

(i) There exists $\epsilon_0 > 0$ *such that for any* $0 < \epsilon < \epsilon_0$*,*

$$\text{pr}\left\{\max_{\mathcal{J}} |\phi_{n\mathcal{J}}^{Smooth} - \phi_{\mathcal{J}}^{Smooth}| > (1+c)\epsilon\right\} \leq (\prod_{r=1}^{R} p_r)(1+\omega)\zeta_n(\epsilon).$$

(ii) If in addition Condition (T1) holds and $d_n \geq |\mathcal{S}_1|$*, there exists a positive constant* $\epsilon_1 = \min\{\epsilon_0, \delta_{\mathcal{S}_1}^{Smooth}/(2(1+c))\}$ *such that*

$$\text{pr}\left\{\mathcal{D} \subseteq \widehat{\mathcal{S}}_{\phi}^{Smooth}(d_n)\right\} \geq 1 - (\prod_{r=1}^{R} p_r)(1+\omega)\zeta_n(\epsilon_1).$$

The first conclusion in Theorem 4.1 implies that, if $\phi_{n\mathcal{J}}$ uniformly converges to $\phi_{\mathcal{J}}$ (i.e., if Condition (T2) holds), its STS version $\phi_{n\mathcal{J}}^{\text{Smooth}}$ uniformly converges to $\phi_{\mathcal{J}}^{\text{Smooth}}$ at the same rate. Most, if not all, existing marginal screening methods satisfy a certain form of Condition (T2), including the three discussed in Section 3.4. Hence, $\phi_{n\mathcal{J}}^{\text{Smooth}}$ in general are very good approximations of $\phi_{\mathcal{J}}^{\text{Smooth}}$, and provide the same rank of the variables as that of $\phi_{\mathcal{J}}^{\text{Smooth}}$.

The second conclusion in Theorem 4.1 indicates that, under the additional Condition (T1), STS enjoys the SURE screening property. Condition (T1) requires $\phi_{\mathcal{J}}^{\text{Smooth}}$ to produce a good ranking of the variables; the important variables have larger $\phi_{\mathcal{J}}^{\text{Smooth}}$ than the unimportant ones on the population level. Then by the first conclusion, the estimates $\phi_{n\mathcal{J}}^{\text{Smooth}}$ should be able to accurately detect the important variables. In this sense, Condition (T1) is similar to Condition (V1) for vector methods. In what follows, we further show that when the true signal is smooth, (T1) is a natural generalization of (V1) to tensor data. First, we rewrite Condition (V1) for tensor data. With a little abuse of terminology, we still refer to the tensor version as Condition (V1).

Condition (V1). *There exists* $\mathcal{S}$ *such that* $\mathcal{D} \subseteq \mathcal{S}$ *and* $\delta_{\mathcal{S}} = \min_{\mathcal{J}\in\mathcal{S}}\{\phi_{\mathcal{J}}\} - \max_{\mathcal{J}\in\mathcal{S}^c}\{\phi_{\mathcal{J}}\} > 0$.

Next, we consider a smoothness assumption. Let $\Phi \in \mathbb{R}^{p_1\times\cdots\times p_R}$ denote the tensor with its $\mathcal{J}$th element being the screening utility $\phi_{\mathcal{J}}$. We make the following assumption:

Assumption (A1). *Assume that there exist J_0 non-overlapping blocks $\{\mathcal{B}_1, \ldots, \mathcal{B}_{J_0}\}$, such that each block contains the indices of a subarray of Φ. The indices of all the other variables that do not fall into any block are in set $\mathcal{B}_{J_0+1}$. Assume that $\mathcal{D} \subseteq \cup_{j=1}^{J_0} \mathcal{B}_j$ and there is a vector $\mathbf{v} \in \mathbb{R}^{J_0}$ such that the true signal Φ can be written as*

$$\Phi = \sum_{j=1}^{J_0} v_j 1_{\mathcal{B}_j},$$

where $\mathbf{v} > 0$ and $1_{\mathcal{B}}$ is the indicator function of set $\mathcal{B}$.

Assumption (A1) is an ideal case where the true signals are piece-wise constant. It can be seen as an approximation to the practical case where the true signals slightly fluctuate within each region. A similar assumption has been used by [44] to study the theoretical properties of the fused lasso.

For a set $\mathcal{S}$ that satisfies Condition (V1), we consider the set $\mathcal{S}_1 = \mathcal{S} \cup \{\cup_{\mathcal{J} \in \mathcal{S}} \Omega_{\mathcal{J}}\}$. The set $\mathcal{S}_1$ contains all the elements in $\mathcal{S}$ and their neighbors. We have the following lemma.

Lemma 4.1. *Under Condition (V1) and Assumption (A1), for any non-negative c, we have $\mathcal{S}_1$ satisfies Condition (T1) and $\delta_{\mathcal{S}_1}^{Smooth} > \min\{1, c/\omega\}\delta_{\mathcal{S}}$. Moreover, $|\mathcal{S}_1| \leq 3^{R-1}(|\mathcal{S}| + 2J_0)$.*

It can be seen that Conditions (V1) & (A1) together imply Condition (T1). Hence, Condition (T1) is very intuitive. Moreover, we generally hope that gap $\delta_{\mathcal{S}_1}^{\text{Smooth}}$ is large. Lemma 4.1 suggests that, when c is considerably smaller than ω, the lower bound for $\delta_{\mathcal{S}_1}^{\text{Smooth}}$ is also much smaller than $\delta_{\mathcal{S}}$. Hence, it makes sense to only consider c comparable to ω, which is in accordance to our proposal of $c = \omega/2$ or $c = \omega$. In the meantime, since R is a fixed constant and $J_0 \leq |\mathcal{S}|$, $|\mathcal{S}_1| \leq 3^{R-1}(|\mathcal{S}| + 2J_0)$ indicates that $|\mathcal{S}_1|$ is at the same order of $|\mathcal{S}|$. Consequently, STS can handle the same level of sparsity as the marginal screening utility it is combined with.

Finally, we note an important difference between screening and hypothesis testing, although many screening utilities are historically test statistics, such as the t-statistic we discussed. In screening, the refined analysis will be performed exclusively on the reduced data. Consequently, if an important variable is falsely removed by screening, it is impossible to recover it in the final modeling. This is why we need the SURE screening property to ensure that there is no false negative. On the other hand, if an unimportant variable survives the screening step, a sparse method in the second step can still identify it. Screening is tolerant of false positives in this sense.

In contrast, in multiple testing problems, it is essential to control the false discovery rate. For example, [41] proposed to use a smoothing procedure to detect clusters. In particular, a bias adjustment is developed to prevent the smoothing from inflating the false discovery rate. No such adjustment is needed in STS because STS is concerned with false negatives rather than false positives.

4.2. Theoretical properties for STS-t, STS-GLM and STS-DC

In what follows, we study the theoretical properties for STS-t, STS-GLM, and STS-DC. For all the three examples, our study verifies that $\phi_{n\mathcal{J}}^{\text{Smooth}}$ uniformly converges to $\phi_{\mathcal{J}}^{\text{Smooth}}$ at the same rate as $\phi_{n\mathcal{J}}$ converges to $\phi_{\mathcal{J}}$ by showing that they have the same order in their probability bounds. For each method, we also replace Condition (T2) with suitable lower-level conditions.

4.2.1. The STS-t screening

According to Theorem 4.1, the SURE screening property of STS-t depends on Conditions (T1) and (T2). We introduce the following condition that guarantees Condition (T2). To apply STS-t screening, we assume the model in (3.4).

Condition (T3). *Within Class $Y = k$, $\epsilon_{k\mathcal{J}}$ is sub-Gaussian with variance proxy $\sigma_{k\mathcal{J}}^2$. Also, we assume that $\mu_{1\mathcal{J}} - \mu_{2\mathcal{J}}$ is bounded uniformly; $\sigma_{1\mathcal{J}}^2$ and $\sigma_{2\mathcal{J}}^2$ are bounded away from 0 uniformly.*

Straightforward proof shows that Condition (T3) implies Condition (T2); see Appendix A.3 for the proof. Condition (T3) is a widely used assumption in high-dimensional statistics. Hence, Condition (T2) is indeed very mild. As a consequence of Theorem 4.1, we have the following result concerning the SURE screening property of STS-t.

Corollary 4.1. *Suppose that Condition (T3) holds. We have the following conclusions.*

(i) There exists $\epsilon_0 > 0$ such that for any $0 < \epsilon < \epsilon_0$,

$$\begin{aligned}&\text{pr}\left\{\max_{\mathcal{J}} |\phi_{n\mathcal{J}}^{t.STS} - \phi_{\mathcal{J}}^{t.STS}| > (1+c)\epsilon\right\}\\ \le&(1+\omega)(\prod_{r=1}^{R} p_r)\{\gamma_1 \exp(-\gamma_2 n\epsilon^2) + \gamma_3 \exp(-\gamma_4 n)\},\end{aligned}$$

for some positive constants γ_1, γ_2, γ_3 and γ_4.

(ii) If in addition Condition (T1) holds, $d_n \ge |\mathcal{S}_1|$ and $\delta_{\mathcal{S}_1}^{Smooth} \gg (\sum_{r=1}^{R} \log(p_r)/n)^{1/2}$, then STS-t enjoys the SURE screening property with a probability tending to 1.

Corollary 4.1 implies that $\phi_{n\mathcal{J}}^{\text{t.STS}}$ converges to its population counterpart at the same rate of $\phi_{n\mathcal{J}}^{t}$. Moreover, STS-t enjoys the SURE screening property even when the dimension of each mode of $\mathbf{X}$ grows at an exponential rate of n. Thus, STS-t is suitable for very high-dimensional tensor datasets.

4.2.2. The STS-GLM screening

For STS-GLM, we study the SURE screening property for two most important generalized linear models, the linear regression model and the logistic regression model. We use the following condition.

Condition (T4). *Both* $\mathbf{X}$ *and* Y *satisfy the sub-exponential tail probability uniformly. That is, there exists a positive constant* s_0 *such that, for all* $0 \leq s \leq 2s_0$,

$$\max_{\mathcal{J}} \mathrm{E}\{\exp\left(sX_{\mathcal{J}}^2\right)\} < \infty \text{ and } \mathrm{E}\{\exp\left(sY^2\right)\} < \infty.$$

Condition (T4) is a mild condition that implies Condition (T2). Detailed proof can be seen in Appendix A.4. Similar conditions have been used in [13] to guarantee the SURE screening property of MMLE for vector data. Under Condition (T4), we establish the SURE screening property of STS-GLM using Theorem 4.1.

Corollary 4.2. *Suppose that Condition (T4) holds. We have the following conclusions.*

(i) For logistic regression, there exist some positive constants α *and* ϵ_0 *such that for any* $0 < \epsilon < \epsilon_0$,

$$\begin{aligned}&\mathrm{pr}\left\{\max_{\mathcal{J}} |\phi_{n\mathcal{J}}^{GLM.STS} - \phi_{\mathcal{J}}^{GLM.STS}| > (1+c)\epsilon\right\}\\ \leq&(1+\omega)(\prod_{r=1}^{R} p_r)\{\gamma_1 \exp(-\gamma_2 n^{\alpha/(\alpha+2)}\epsilon^2) + \gamma_3 n \exp(-\gamma_4 n^{\alpha/(\alpha+2)})\},\end{aligned}$$

for some positive constants γ_1, γ_2, γ_3 *and* γ_4. *Moreover, if Condition (T1) holds,* $d_n \geq |\mathcal{S}_1|$ *and* $\delta_{\mathcal{S}_1}^{Smooth} \gg (\sum_{r=1}^{R} \log(p_r)/n^{\alpha/(\alpha+2)})^{1/2}$, *then STS-GLM enjoys the SURE screening property with a probability tending to 1.*

(ii) For linear models, there exist some positive constants α *and* ϵ_0 *such that for any* $0 < \epsilon < \epsilon_0$,

$$\begin{aligned}&\mathrm{pr}\left\{\max_{\mathcal{J}} |\phi_{n\mathcal{J}}^{GLM.STS} - \phi_{\mathcal{J}}^{GLM.STS}| > (1+c)\epsilon\right\}\\ \leq&(1+\omega)(\prod_{r=1}^{R} p_r)\{\gamma_1 \exp(-\gamma_2 n^{\alpha/A}\epsilon^2)\},\end{aligned}$$

for some positive constants γ_1, γ_2 *and* $A = \max(\alpha+4, 3\alpha+2)$. *Moreover, if Condition (T1) holds,* $d_n \geq |\mathcal{S}_1|$ *and* $\delta_{\mathcal{S}_1}^{Smooth} \gg (\sum_{r=1}^{R} \log(p_r)/n^{\alpha/A})^{1/2}$, *then STS-GLM enjoys the SURE screening property with a probability tending to 1.*

Corollary 4.2 suggests that STS-GLM can also handle the tensor data with very high dimensionality along each mode.

4.2.3. The STS-DC screening

To study the theoretical properties of STS-DC, we again consider Condition (T4). Condition (T4) was used in [25] to guarantee the SURE screening property of the distance correlation screening method for vector data. In Appendix A.4, we showed that Condition (T4) implies Condition (T2). Therefore, we establish the SURE screening property of STS-DC in the following corollary.

Corollary 4.3. *Suppose that Condition (T4) holds. We have the following conclusions.*

(i) There exists $\epsilon_0 > 0$ *such that for any* $0 < \epsilon < \epsilon_0$ *and for any* $0 < v < 1/2$,

$$\begin{aligned}&\mathrm{pr}\left\{\max_{\mathcal{J}} |\phi_{n\mathcal{J}}^{DC.STS} - \phi_{\mathcal{J}}^{DC.STS}| > (1+c)\epsilon\right\}\\ &\leq (1+\omega)(\prod_{r=1}^{R} p_r)\{\gamma_1 \exp\left(-\gamma_2 n^{1-2v}\epsilon^2\right) + \gamma_3 n \exp(-\gamma_4 n^v)\},\end{aligned}$$

for some positive constants γ_1, γ_2, γ_3 *and* γ_4.

(ii) If in addition Condition (T1) holds, $d_n \geq |\mathcal{S}_1|$ *and* $\delta_{\mathcal{S}_1}^{Smooth} \gg (\sum_{r=1}^{R} \log(p_r)/n^{1/3})^{1/2}$ *where we let* $v = 1/3$*, then STS-DC enjoys the SURE screening property with a probability tending to 1.*

By Corollaries 4.1, 4.2 and 4.3, all the three STS screening methods enjoy the SURE screening property when the dimension of each mode of the tensor grows at an exponential rate of the sample size. We also note that their convergence rates are identical to those of their marginal counterparts. Hence, the STS procedure reserves the nice theoretical properties of their marginal counterparts while taking advantage of the tensor structure.

4.3. Structure of the screened data

As suggested by a referee, we further examine the impact of screening on modeling. We are interested in whether the screened data follow the same type of model as the original data. For such study, although screening can be performed in a model-free fashion, we consider two popular tensor models that could be fitted after screening, the tensor discriminant analysis (TDA) model [42], and the generalized linear tensor regression model [59].

There are two possible results of STS. On one hand, if we simply apply STS, the predictor may no longer be a tensor and cannot be modeled by tensor models. On the other hand, if we combine STS with the augmentation introduced in Section 3.3, we will end up with a sub-tensor of $\mathbf{X}$. For simplicity, we pay more attention to the latter case, as it is more relevant for tensor data analysis. The former case is briefly discussed afterwards.

Denote $\mathcal{V}$ as a subset of indices such that $\mathbf{X}_{\mathcal{V}}$ is also a tensor. In the context of screening, $\mathcal{V}$ could be the target set for STS with augmentation. Note that, since $\mathbf{X}_{\mathcal{V}}$ is a tensor, there exist $\mathcal{V}_r \subset \{1, \ldots, p_r\}, r = 1, \ldots, R$ such that $\mathcal{V} = \mathcal{V}_1 \times \cdots \times \mathcal{V}_R$, where $\times$ denotes the Cartesian product.

For the TDA model, $Y \in \{1, \ldots, K\}$ is the class label. We assume that

$$\mathrm{pr}(Y = k) = \pi_k, \quad \mathbf{X} \mid (Y = k) \sim TN\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R\right),$$

where $0 < \pi_k < 1, \sum_k \pi_k = 1$ are the prior probabilities for Class k, $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ is the within-class mean, and $\boldsymbol{\Sigma}_r, r = 1, \ldots, R$ are positive definite covariance matrices.

Note that STS-t, STS-GLM and STS-DC are all applicable under this model. For example, STS-t is suitable because the TDA model satisfies Condition (T3), which eventually leads to the SURE screening property. To see why Condition (T3) is true, we note that the TDA model can be equivalently written as (3.4) with the additional assumption that $\boldsymbol{\epsilon}_k \sim TN(0, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R)$. It follows that, for each $\mathcal{J}$, $\epsilon_{k\mathcal{J}} \sim N(0, \sigma^2_{k\mathcal{J}})$ with $\sigma^2_{k\mathcal{J}} = \prod_{r=1}^R \sigma_{r,j_r j_r}$. Hence, $\epsilon_{k\mathcal{J}}$ is sub-Gaussian with variance proxy $\sigma^2_{k\mathcal{J}}$, as required by Condition (T3).

Under the TDA model, we have the following lemma for the sub-tensor $\mathbf{X}_\mathcal{V}$.

Lemma 4.2. *Under the TDA model, for any index set $\mathcal{V}$, we have*

$$\mathbf{X}_\mathcal{V} \mid (Y = k) \sim TN\left(\boldsymbol{\mu}_{k,\mathcal{V}}, \boldsymbol{\Sigma}_{1,\mathcal{V}_1}, \ldots, \boldsymbol{\Sigma}_{R,\mathcal{V}_R}\right),$$

where $\boldsymbol{\Sigma}_{r,\mathcal{V}_r}$ is the sub-matrix of $\boldsymbol{\Sigma}_r$ containing elements with indices in $\mathcal{V}_r \times \mathcal{V}_r$.

Lemma 4.2 implies that any sub-tensor $\mathbf{X}_\mathcal{V}$ preserves the TDA model. Therefore, if $(\mathbf{X}, Y)$ follows the TDA model, we can still fit a TDA model after STS with data augmentation.

Meanwhile, the generalized linear tensor regression model considers a univariate response Y that could be continuous or discrete. It assumes that

$$g(\mu) = \beta_0 + \langle \mathbf{B}, \mathbf{X} \rangle \tag{4.2}$$

where $g(\cdot)$ is the link function, $\mu = \mathrm{E}(Y \mid \mathbf{X})$, β_0 is the intercept and $\mathbf{B}$ has rank H. The inner product between two tensors is defined as $\langle \mathbf{B}, \mathbf{X} \rangle = \sum_\mathcal{J} B_\mathcal{J} X_\mathcal{J}$. We also refer to this model as the rank-H generalized linear tensor regression model to highlight the low-rank structure. Recall that, we assume that $\mathbf{B}$ is sparse and is only nonzero over a set $\mathcal{D}$. We have the following lemma for $\mathbf{X}_\mathcal{V}$.

Lemma 4.3. *Under the rank-H generalized linear tensor regression model, if $\mathcal{D} \subseteq \mathcal{V}$, we have*

$$g(\mu) = \beta_0 + \langle \mathbf{B}_\mathcal{V}, \mathbf{X}_\mathcal{V} \rangle, \tag{4.3}$$

where $\mathbf{B}_\mathcal{V}$ also has rank H.

Lemma 4.3 indicates that, if a sub-tensor $\mathbf{X}_\mathcal{V}$ contains all the important elements, then it is connected to Y with the rank-H generalized linear tensor regression model. In our context, take $\mathcal{V}$ to be the target set of STS with data augmentation. If $\mathcal{D} \subseteq \mathcal{V}$, the rank-$H$ generalized linear tensor regression is preserved after STS with data augmentation. The assumption $\mathcal{D} \subseteq \mathcal{V}$ is reasonable because we expect STS to enjoy the SURE screening property when we apply it.

Finally, we point out that STS alone (without data augmentation) also preserves some important properties of the original model. For example, if $(\mathbf{X}, Y)$ follows the TDA model, then it can be shown that the screened data follows the linear discriminant analysis model, which is the counterpart for the TDA model on vector data. Similarly, if $(\mathbf{X}, Y)$ follows the generalized linear regression tensor model, the screened data follow the generalized linear model as long as all the important variables are kept.

5. Numerical studies

5.1. Simulations

We present the numerical performance of STS. In all simulations, we generate independent training, validation, and testing sets. The training set and validation set contain n observations to be specified, while the testing set contains 10000 observations. All matrix predictors ($R = 2$) have dimension 64×64. Three-way tensor predictors ($R = 3$) have dimension $30 \times 36 \times 30$.

We study classification models and regression models under both smooth and non-smooth settings. We first introduce the classification models with smoothness. In each model, there are 2 classes and 75 observations within each class. The predictor $\mathbf{X}$ is generated from the tensor discriminant model with $K = 2$: $\mathbf{X} \mid (Y = k) \sim TN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R)$ where $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots p_R}$, $\boldsymbol{\Sigma}_r \in \mathbb{R}^{p_r \times p_r}, r = 1, \ldots, R$ and $\pi_k = \mathrm{pr}(Y = k)$, $k = 1, 2$. By [42], the best classifier under this model is

$$\widehat{Y} = 1\left\{\log\left(\pi_2/\pi_1\right) + \langle \mathbf{B}, \mathbf{X} - \left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right)/2\rangle > 0\right\} + 1,$$

where $\mathbf{B} = [\![\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_R^{-1}]\!]$. Therefore, to ensure sparsity on the population level, we specify sparse $\mathbf{B}$, along with $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R$. Then the mean tensors are set to be $\boldsymbol{\mu}_1 = 0$ and $\boldsymbol{\mu}_2 = [\![\mathbf{B}; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R]\!]$. The active set $\mathcal{D}$ of $\mathbf{B}$ is defined as

$$\mathcal{D} = \{\mathcal{J} : b_{\mathcal{J}} \neq 0\}.$$

In our model settings we only specify elements of $\mathbf{B}$ in $\mathcal{D}$. It is always assumed that $b_{\mathcal{J}} = 0$ for any $\mathcal{J} \in \mathcal{D}^c$. For an index set $\mathcal{A}$, we use the notation $\mathbf{B}_{\mathcal{A}} = a$ to denote that $b_{\mathcal{J}} = a$ for any $\mathcal{J} \in \mathcal{A}$ where a is a number. For ease of presentation, we define two sets of integers, $L_1 = 31 : 34$ and $L_2 = 21 : 23$. For a covariance matrix $\boldsymbol{\Sigma}$, we use the following notations. Let $\boldsymbol{\Sigma} = \mathrm{AR}(\rho)$ denote that $\boldsymbol{\Sigma}$ is autoregressive, i.e., $\sigma_{ij} = \rho^{|i-j|}$. Let $\boldsymbol{\Sigma} = \mathrm{CS}(\rho)$ denote that $\boldsymbol{\Sigma}$ has the compound symmetry structure, i.e., $\sigma_{ij} = \rho$, $i \neq j$. In all simulations we consider $\pi_1 = \pi_2 = 1/2$. We consider the following three smooth classification models.

Model 1: $K = 2, R = 2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_{64}$, $\mathcal{D} = \{(i, j) : i, j \in L_1\}$, $\mathbf{B}_{\mathcal{D}} = 0.6$.

Model 2: $K = 2, R = 2$, $\boldsymbol{\Sigma}_1 = \mathrm{AR}(0.3)$, $\boldsymbol{\Sigma}_2 = \mathbf{I}_{64}$, $\mathcal{D}$ is a diamond area with 25 variables where the four vertexes are located at $(30, 20)$, $(36, 20)$, $(33, 17)$, $(33, 23)$ and $\mathbf{B}_{\mathcal{D}} = 0.4$.

Model 3: $K = 2, R = 3$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_{30}$, $\boldsymbol{\Sigma}_2 = \mathrm{CS}(0.2)$, $\boldsymbol{\Sigma}_3 = \mathrm{CS}(0.2)$, $\mathcal{D} = \{(i, j, k) : i, j, k \in L_2\}$, $\mathbf{B}_{\mathcal{D}} = 0.4$.

For regression models, we set $n = 200$. Define $\mathbf{B}$ as the regression coefficient. The active set $\mathcal{D}$ is defined as

$$\mathcal{D} = \{\mathcal{J} : b_{\mathcal{J}} \neq 0\}.$$

We consider the following four models.

Model 4: $Y = \langle \mathbf{B}, \mathbf{X} \rangle + \epsilon$ where $R = 2$, $\mathbf{X} \sim MN(0, \mathbf{I}_{64}, \mathbf{I}_{64})$, $\epsilon \sim N(0,1)$, $\mathcal{D} = \{(i,j) : i,j \in L_1\}$ and $\mathbf{B}_{\mathcal{D}} = 2$.

Model 5: $Y = 2\sum_{\mathcal{J}\in\mathcal{D}_1} X_{\mathcal{J}} + 0.8\sum_{\mathcal{J}\in\mathcal{D}_2} X_{\mathcal{J}}^3 + \epsilon$, where $R = 2$, $\mathbf{X} \sim MN(0, \mathrm{AR}(0.5), \mathbf{I}_{64})$, $\epsilon \sim N(0,1)$, $\mathcal{D} = \mathcal{D}_1 \bigcup \mathcal{D}_2$, $\mathcal{D}_1 = \{(i,j) : i,j \in L_1\}$, $\mathcal{D}_2$ is a triangle area with 9 variables and the three vertexes are located at $(50,50)$, $(52,48)$, $(52,52)$.

Model 6: $Y = \langle \mathbf{B}, \mathbf{X} \rangle + \epsilon$ where $R = 3$, $\mathbf{X} \sim MN(0, \mathrm{AR}(0.6), \mathbf{I}_{36}, \mathrm{CS}(0.3))$, $\epsilon \sim N(0,1)$, $\mathcal{D} = \{(i,j,k) : i,j,k \in L_2\}$ and $\mathbf{B}_{\mathcal{D}} = 0.8$.

Model 7: $Y \sim \mathrm{Poisson}(\exp\{\langle \mathbf{B}, \mathbf{X} \rangle\})$ where $R = 2$, $\mathbf{X} \sim MN(0, \mathbf{I}_{64}, \mathbf{I}_{64})$, $\mathcal{D} = \{(i,j) : i,j \in L_2\}$ and $\mathbf{B}_{\mathcal{D}} = 0.3$.

To study the robustness of our proposed method, we also include the following four models with model misspecification. Models 8 & 9 are classification models and Models 10 & 11 are regression models. In Models 8 & 10, the important variables are chosen completely at random and there is no smoothness structure. In Models 9 & 11, the active set is a vertical line segment and only has smoothness along one direction. In other words, in Models 8 & 10 the smoothness assumption does not make sense at all, but in Models 9 & 11 the smoothness assumption is partially true.

Model 8: $(\mathbf{X}, Y)$ follows the tensor discriminant analysis model with $K = 2$, $R = 2$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_{64}$, the active set $\mathcal{D}$ consists of 16 randomly chosen predictors, $\mathbf{B}_{\mathcal{D}} = 1$. There are 75 observations within each class.

Model 9: $(\mathbf{X}, Y)$ follows the tensor discriminant analysis model with $K = 2$, $R = 2$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_{64}$, $\mathcal{D} = \{(i,j) : 21 \le i \le 30, j = 10\}$, $\mathbf{B}_{\mathcal{D}} = 0.7$. There are 75 observations within each class.

Model 10: $n = 200$, $Y = \langle \mathbf{B}, \mathbf{X} \rangle + \epsilon$ where $R = 2$, $\mathbf{X} \sim MN(0, \mathbf{I}_{64}, \mathbf{I}_{64})$, $\epsilon \sim N(0,1)$, the active set $\mathcal{D}$ consists of 9 randomly chosen predictors, $\mathbf{B}_{\mathcal{D}} = 2$.

Model 11: $n = 200$, $Y = \langle \mathbf{B}, \mathbf{X} \rangle + \epsilon$ where $R = 2$, $\mathbf{X} \sim MN(0, \mathbf{I}_{64}, \mathbf{I}_{64})$, $\epsilon \sim N(0,1)$, $\mathcal{D} = \{(i,j) : 21 \le i \le 32, j = 10\}$, $\mathbf{B}_{\mathcal{D}} = 0.8$.

For classification models (Models 1–3, 8–9), we consider STS-t, STS-GLM and STS-DC. For regression models (Models 4–7, 10–11), we use STS-GLM and STS-DC. We choose c from $\{\omega/2, \omega\}$, that is, $\{4, 8\}$ for matrix models and $\{13, 26\}$ for three-way tensor models. We compare these methods with the corresponding marginal screening methods.

The minimum numbers of variables needed to recover all active predictors are reported in Table 1. The reported numbers are medians from 500 replicates and their standard errors computed from bootstrap. A closer number to $|\mathcal{D}|$ indicates a better screening technique. It can be seen that the STS methods are uniformly more efficient than their marginal counterparts in identifying the important variables in Models 1–7. This demonstrates the benefits of leveraging the tensor structure for better screening when the smoothness structure is present. In these models, often both $c = \omega/2$ and $c = \omega$ give good results. For Models 8 & 10 where there is completely no smoothness structure, STS is worse than marginal screening due to model misspecification, although $c = \omega/2$ is significantly better than $c = \omega$. Thus, when the smoothness assumption does not hold at all, STS is not recommended. However, for Models 9 & 11 where there is partial smoothness structure along one direction, STS again outperforms marginal screening with

TABLE 1

Minimum numbers needed to recover all active predictors for all models. Bootstrap is used to calculate the standard errors which are reported in the parentheses. In the second row, $\omega/2$ and ω are the values for c. Since Models 4–7 and 10–11 are regression models, STS-t is not applicable.

Model	$\lvert\mathcal{D}\rvert$	STS-t			STS-DC			STS-GLM		
		Marginal	$\omega/2$	ω	Marginal	$\omega/2$	ω	Marginal	$\omega/2$	ω
1	16	238.5	17	20	326	18	21	265	17	20
		(14.5)	(0)	(0.3)	(15.2)	(0.3)	(0.2)	(17.7)	(0.3)	(0.5)
2	25	571	32	34	724.5	33	34	597.5	33	35
		(34.4)	(0.4)	(0.3)	(40.8)	(0.2)	(0.2)	(35.2)	(0.4)	(0.4)
3	27	340.5	61	73	411	44	48	458.5	63	74
		(21.0)	(1.4)	(1.4)	(22.7)	(0.5)	(0.4)	(26.3)	(1.3)	(1.6)
4	16	-	-	-	536	19	22	312	17	21
		-	-	-	(36.0)	(0.2)	(0.2)	(21.1)	(0.4)	(0.5)
5	25	-	-	-	1437.5	133	129.5	986.5	90.5	88
		-	-	-	(86)	(9.8)	(10)	(80)	(8.5)	(5.3)
6	27	-	-	-	353	54	57	201	93	99
		-	-	-	(20.8)	(1.1)	(1.2)	(13.3)	(1.9)	(2.0)
7	9	-	-	-	416	11	13	375	11	13
		-	-	-	(21.8)	(0.2)	(0.4)	(29.5)	(0.1)	(0.5)
8	16	16	83.5	671.5	16	79	176	16	108	734.5
		(0)	(3.4)	(17.2)	(0)	(1.8)	(3.9)	(0.3)	(3.9)	(24.4)
9	10	36	25	45	47	26	31	51	26	48
		(2.1)	(0.6)	(1.5)	(2.4)	(0.4)	(0.5)	(3.0)	(0.6)	(2.2)
10	9	-	-	-	18	95.5	446.5	13	182	1161
		-	-	-	(1.2)	(5.2)	(23.2)	(0.5)	(7.7)	(34.5)
11	12	-	-	-	179.5	35	47	92	37	88
		-	-	-	(12.4)	(0.8)	(1.9)	(6.7)	(1.5)	(5.9)

both choices of c, but $c = \omega/2$ is better than $c = \omega$. Therefore, if it is believed that some level of smoothness exists, $c = \omega/2$ is a robust choice to be used in STS.

To further demonstrate that STS can improve the analysis accuracy, we apply different classification and regression methods on the screened data. For all models, if screening is applied, we let $d_n = \lceil n/\log n \rceil$. For classification, we include two tensor-based methods, covariate-adjusted tensor classification (CATCH, [42]) and tensor regression based on CP decomposition (CP-GLM, [59]), and two vector-based methods, ℓ_1-penalized generalized linear model (ℓ_1-GLM, [14]) and ℓ_1-penalized Fisher's discriminant analysis (ℓ_1-FDA, [54]). For regression, we include CP-GLM and ℓ_1-GLM. For the implementation of CATCH, ℓ_1-GLM, ℓ_1-FDA, we used the `R` packages `catch`, `glmnet` and `penalizedLDA`. For the implementation of CP-GLM, we used the `MATLAB` toolbox `TensorReg` downloaded

from `https://hua-zhou.github.io/TensorReg`.

STS based on the three statistical utilities give similar patterns of performance. For the sake of space, here we only present the results using regression-based screening, that is, logistic regression for Models 1–3 & 8–9, linear regression for Models 4–6 & 10–11 and Poisson regression for Model 7. The simulation results using t-statistic and distance correlation for classification models can be found in Appendix B.

We report the model fitting results of 100 replicates in Table 2. For classification models, we report the classification errors and their standard errors, while for regression models we report the root mean square error (RMSE), which is defined as

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{Y}^i - Y^i)^2}{n}}.$$

We further report the true positive rate (TPR) and the false positive rate (FPR) for classification models in Table 3. Let $\widehat{\mathcal{D}}$ be the set of active predictors in the final model. The TPR and FPR are defined as

$$\text{TPR} = \frac{|\widehat{\mathcal{D}} \cap \mathcal{D}|}{|\mathcal{D}|}, \quad \text{FPR} = \frac{\left|\widehat{\mathcal{D}} \cap \mathcal{D}^c\right|}{|\mathcal{D}^c|}.$$

According to Table 2, the comparison among different c and the original data suggests that smoothing screening statistics is critical to tensor screening when smooth structure exists (i.e., Models 1–7). Even though the classification errors and RMSE are reduced for some methods with marginal screening, the minimum error is obtained at $c = \omega/2$ or ω, and the decrease is significant. In Table 3, STS successfully increases the true positive rate while reducing the false positive rate. Higher TPR and lower FPR are achieved at $c = \omega/2$ or ω compared to $c = 0$. This coincides with the changing trend in classification errors, which further shows the importance of smoothing screening statistics. Moreover, in several cases such as CP-GLM in Models 3, 5 & 6, the standard errors are lowered after applying screening with or without STS. This indicates that screening tends to make the analysis more stable. For non-smooth models (Models 8 & 10), STS has lower accuracy and worse variable selection results compared to marginal screening, which is a consequence of model misspecification. However, STS is again better than marginal screening combined with most model-fitting methods when partial smoothness exists (Models 9 & 11).

5.2. Real data analysis

Electroencephalography (EEG) is used to record the electrical activity of human brains and to detect the associated brain disorders. The dataset we use arises from a study that analyzes the relationship between EEG and alcoholism, available at `http://kdd.ics.uci.edu/databases/eeg/eeg.data.html`. There are

TABLE 2

Classification errors (%) for Models 1-3, 8-9, RMSE for Models 4-7, 10-11 and their standard errors are reported. The Bayes errors for Models 1-3, 8-9 are 11.48%, 10.74%, 7.26%, 2.27% and 13.44% respectively. STS-GLM is used for screening. For each model, the column $\mathbf{X}$ *corresponds to the results on the original dataset; the other three columns are results on data with different screening methods: marginal screening, and STS with two choices of c.*

	Model 1				Model 2				Model 3			
	$\mathbf{X}$	0	$\omega/2$	ω	$\mathbf{X}$	0	$\omega/2$	ω	$\mathbf{X}$	0	$\omega/2$	ω
CATCH	20.04	19.24	15.71	15.00	18.11	17.48	13.90	13.66	13.17	11.74	9.55	9.60
	(0.27)	(0.27)	(0.16)	(0.16)	(0.18)	(0.18)	(0.14)	(0.12)	(0.15)	(0.15)	(0.09)	(0.09)
CP-GLM	21.87	21.74	21.50	21.50	20.95	22.14	19.52	19.22	19.81	14.15	12.85	12.98
	(0.70)	(0.35)	(0.30)	(0.30)	(0.38)	(0.38)	(0.36)	(0.27)	(1.26)	(0.35)	(0.30)	(0.38)
ℓ_1-GLM	23.10	22.89	17.63	16.27	20.15	19.70	15.73	15.49	15.69	15.28	12.55	12.82
	(0.32)	(0.32)	(0.22)	(0.15)	(0.19)	(0.22)	(0.14)	(0.14)	(0.19)	(0.23)	(0.14)	(0.14)
ℓ_1-FDA	31.95	21.34	15.01	13.90	18.77	16.37	12.37	12.30	16.08	12.79	10.18	10.64
	(0.36)	(0.28)	(0.11)	(0.10)	(0.24)	(0.18)	(0.09)	(0.07)	(0.30)	(0.18)	(0.08)	(0.09)
	Model 4				Model 5				Model 6			
CP-GLM	1.69	1.04	1.04	1.04	11.60	7.93	7.37	7.33	2.70	1.04	1.03	1.03
	(2.32)	(0.02)	(0.02)	(0.02)	(2.67)	(0.73)	(0.38)	(0.38)	(3.56)	(0.02)	(0.02)	(0.02)
ℓ_1-GLM	1.57	4.58	1.10	1.19	9.49	10.04	7.67	7.73	1.56	2.46	2.47	2.71
	(0.14)	(0.93)	(0.03)	(0.35)	(0.77)	(1.01)	(0.47)	(0.43)	(0.15)	(0.63)	(0.70)	(0.74)
	Model 7				Model 8				Model 9			
CATCH	-	-	-	-	3.30	3.26	4.54	14.33	18.98	18.96	17.09	16.75
	-	-	-	-	(0.06)	(0.06)	(0.12)	(0.36)	(0.20)	(0.19)	(0.17)	(0.17)
CP-GLM	3.05	3.03	2.99	2.99	26.17	20.91	21.34	27.08	23.53	22.16	19.27	19.27
	(0.20)	(0.22)	(0.18)	(0.17)	(0.56)	(0.37)	(0.39)	(0.50)	(0.59)	(0.41)	(0.35)	(0.34)
ℓ_1-GLM	1.89	1.91	1.62	1.52	4.82	4.58	5.74	22.70	20.97	20.76	18.18	18.50
	(0.12)	(0.11)	(0.12)	(0.12)	(0.10)	(0.10)	(0.12)	(0.61)	(0.25)	(0.25)	(0.19)	(0.19)
ℓ_1-FDA	-	-	-	-	3.18	3.34	4.59	23.06	26.89	20.08	16.82	17.25
	-	-	-	-	(0.04)	(0.05)	(0.10)	(0.63)	(0.47)	(0.22)	(0.15)	(0.16)
	Model 10				Model 11							
CP-GLM	6.30	5.56	5.39	5.63	1.23	1.35	1.22	1.30				
	(0.44)	(0.29)	(0.26)	(0.30)	(0.06)	(0.03)	(0.02)	(0.02)				
ℓ_1-GLM	1.26	1.31	3.52	5.56	1.38	1.60	1.30	1.59				
	(0.06)	(0.53)	(0.89)	(0.37)	(0.01)	(0.03)	(0.02)	(0.03)				

122 subjects, including 77 alcoholic individuals and 45 nonalcoholic individuals. Each subject is exposed to either a single stimulus or to two stimuli that are pictures chosen from [47]. For two stimuli, the two pictures can be either identical or different. Each subject completes 120 trials under different stimuli. In all trials, 64 electrodes are placed on a subject's scalp which are sampled at 256 Hz (3.9-msec epoch) for 1 second and the voltage fluctuates are collected. More information about the collection process can be found in [58]. The same dataset is also analyzed in [23]. They only consider the single stimulus condition and take the average of all the trials under that condition. We use the same part of the data. Thus, we have each predictor being an EEG image of size 256×64

TABLE 3

TPR and FPR (%) comparison for Models 1-3 and 8-9. STS-GLM is used in screening. For each model, the column $\mathbf{X}$ *corresponds to the results on the original dataset; the other three columns are results on data with different screening methods: marginal screening, and STS with two choices of c.*

		$\mathbf{X}$	0	$\omega/2$	ω	$\mathbf{X}$	0	$\omega/2$	ω	$\mathbf{X}$	0	$\omega/2$	ω
		Model 1				Model 2				Model 3			
CATCH	FPR	0.90	0.64	0.55	0.60	0.91	0.40	0.23	0.8	0.12	0.07	0.05	0.05
	S.E.	(0.07)	(0.05)	(0.05)	(0.05)	(0.09)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
	TPR	81.56	87.44	97.06	97.75	68.24	71.04	89.44	89.04	68.33	79.52	91.30	90.52
	S.E.	(1.43)	(1.48)	(0.60)	(0.66)	(1.34)	(1.52)	(1.07)	(0.91)	(1.07)	(1.22)	(0.84)	(0.83)
CP-GLM	FPR	13.20	4.26	1.98	1.76	13.35	3.17	0.62	0.51	7.39	0.63	0.13	0.14
	S.E.	(0.48)	(0.16)	(0.07)	(0.07)	(0.39)	(0.12)	(0.03)	(0.02)	(0.59)	(0.04)	(0.01)	(0.01)
	TPR	96.38	98.63	99.38	99.50	87.76	85.68	94.20	93.40	85.33	99.15	98.78	99.59
	S.E.	(1.71)	(0.62)	(0.21)	(0.28)	(1.06)	(0.78)	(0.53)	(0.63)	(3.04)	(0.39)	(0.41)	(0.20)
ℓ_1-GLM	FPR	0.55	0.24	0.16	0.14	0.48	0.17	0.05	0.05	0.07	0.02	0.01	0.01
	S.E.	(0.05)	(0.01)	(0.01)	(0.01)	(0.05)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
	TPR	67.81	66.13	92.13	93.00	55.28	56.96	76.56	76.96	53.07	51.93	64.63	61.67
	S.E.	(1.38)	(1.29)	(0.94)	(0.71)	(1.00)	(1.08)	(1.22)	(1.13)	(0.89)	(0.89)	(0.81)	(0.74)
ℓ_1-FDA	FPR	33.15	0.40	0.25	0.22	9.26	0.27	0.10	0.11	9.45	0.04	0.02	0.02
	S.E.	(1.49)	(0.01)	(0.01)	(0.01)	(0.48)	(0.01)	(0.00)	(0.00)	(1.23)	(0.00)	(0.00)	(0.00)
	TPR	98.38	72.75	99.00	98.69	94.48	70.76	95.44	94.60	95.59	62.00	77.30	74.04
	S.E.	(0.70)	(1.03)	(0.34)	(0.31)	(0.89)	(0.87)	(0.48)	(0.39)	(1.10)	(0.86)	(0.46)	(0.46)
		Model 8				Model 9							
CATCH	FPR	0.42	0.23	0.19	0.14	0.66	0.38	0.28	0.26				
	S.E.	(0.08)	(0.04)	(0.03)	(0.02)	(0.07)	(0.03)	(0.03)	(0.03)				
	TPR	98.50	98.94	80.81	32.69	90.90	88.90	93.70	92.40				
	S.E.	(0.40)	(0.32)	(0.88)	(0.94)	(1.12)	(1.02)	(0.81)	(0.79)				
CP-GLM	FPR	10.65	4.45	3.75	3.78	9.62	2.63	0.60	0.50				
	S.E.	(0.56)	(0.16)	(0.11)	(0.16)	(0.62)	(0.22)	(0.06)	(0.05)				
	TPR	37.81	60.25	58.06	29.69	92.60	92.50	96.00	93.10				
	S.E.	(1.02)	(1.22)	(1.15)	(0.93)	(1.97)	(0.77)	(0.60)	(0.63)				
ℓ_1-GLM	FPR	0.21	0.08	0.05	0.09	0.44	0.22	0.15	0.12				
	S.E.	(0.03)	(0.01)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.01)				
	TPR	92.44	93.56	77.00	16.75	82.50	83.60	91.00	84.00				
	S.E.	(0.79)	(0.72)	(0.89)	(0.67)	(1.23)	(1.10)	(0.95)	(0.93)				
ℓ_1-FDA	FPR	0.68	0.29	0.26	0.28	11.46	0.44	0.25	0.26				
	S.E.	(0.06)	(0.01)	(0.01)	(0.02)	(0.86)	(0.01)	(0.01)	(0.02)				
	TPR	99.88	99.88	78.50	16.75	95.60	89.10	94.90	86.40				
	S.E.	(0.09)	(0.09)	(0.89)	(0.67)	(1.35)	(0.94)	(0.66)	(0.87)				

and a binary response variable indicating whether the subject is alcoholic or nonalcoholic.

Given the large number of predictors, we assume a sparse model for the prediction of the alcoholic status. For the predictor $\mathbf{X} \in \mathbb{R}^{256\times 64}$, we assume that there exist a few entries X_{ij} that are responsible for the prediction. In the context of the EEG data, we assume that only the measurements of a small

number of electrodes at a few time points are helpful for predicting the alcoholic status. We further assume that the sparsity pattern is smooth, in that X_{ij}'s close to each other tend to be important or unimportant at the same time.

The smoothness of EEG data results from two aspects. First, for each row of an EEG observation, the numbers are the voltages collected over a continuous time period. The voltage will fluctuate when the brain responds to the stimulus and then decay gradually in time. Second, each column of the image represents the voltages collected from 64 different electrodes. Most of these electrodes are entered into the dataset in a way such that two adjacent columns correspond to electrodes positioned symmetrically on the left and right sides of the brain. For example, Column 1 contains measurements from the position FP1, and Column 2 contains those from FP2. FP1 and FP2 are on the same location of the left and right hemispheres [46]. Symmetric locations are functionally similar and thus may resemble each other in the response to stimuli.

Data are standardized before use. The data is randomly split 500 times. In each replicate, we randomly split the data with a $4:1$ ratio into a training set of 97 subjects and a testing set of 25 subjects. For all classification methods available for cross-validation, we set the number of folds to be 5. We choose $d_n = \lceil n/\log(n) \rceil = 21$ and weight c is chosen from the set $\{0, 4, 8\}$ in screening. We use STS based on three different statistical utilities, the t-statistic, logistic regression, and distance correlation. Classification errors on the original data and the screened data for the three methods are reported in Table 4. The results show that screening can either maintain or lower the error rate for most methods. Comparing the results at $c = \omega/2$ or ω with the results at $c = 0$, it can be seen that error rates are further reduced after we add a weight in screening, which supports the application of STS. We also want to mention that, in almost all the analyses, we directly perform screening and/or model fitting on the original dataset of dimension 256×64, with the only exception for CP-GLM without screening. CP-GLM requires the sample size to be no smaller than the dimension of $\mathbf{X}$ on each mode, so when it is applied to the original dataset, we downsize the predictor to 32×32 first and then fit a rank-1 model. This can also be viewed as an advantage of STS. With STS, the dimension of predictor is largely reduced so that CP-GLM can be easily fit without further downsizing, and we observe a uniform decrease in classification errors this way.

To further validate the smoothness assumption among electrodes, we randomly shuffle the columns of the matrix predictors so that it is certain that no smoothness structure exists along this mode. We perform STS and model fitting on the shuffled data. The classification error rates are reported in Table 5. It can be seen that the error rates are generally higher when we change the order of the columns, indicating that the original ordering has some useful information. Moreover, the lowest error rate achieved after shuffling is significantly larger than the error rate without shuffling. Since the lowest error rate is achieved by ℓ_1-GLM in Table 4, we perform a paired t-test for the results before and after shuffling for ℓ_1-GLM at $c = \omega/2$. The p-values under STS-t, STS-GLM, STS-DC are $2.0 \times 10^{-12}, 1.2 \times 10^{-7}, 9.6 \times 10^{-10}$, respectively. Thus, analysis on the original dataset is significantly better, which again provides evidence of

TABLE 4

The means and standard errors of binary classification error rates on EEG dataset. The column $\mathbf{X}$ *corresponds to the classification error rates for models fitted on the original dataset; the other three columns are error rates for models fitted on data, with different screening methods: marginal screening, and STS with two choices of* c.

Screening Method		X		Marginal		$c = \omega/2$		$c = \omega$	
		Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
t-statistic	CATCH	23.66	(0.34)	23.18	(0.31)	21.93	(0.31)	22.19	(0.30)
	CP-GLM	24.66	(0.36)	24.57	(0.34)	22.40	(0.35)	22.72	(0.34)
	ℓ_1-GLM	24.20	(0.32)	23.51	(0.33)	21.45	(0.33)	21.74	(0.34)
	ℓ_1-FDA	26.11	(0.38)	23.86	(0.34)	21.73	(0.32)	22.19	(0.32)
Logistic Regression	CATCH	23.66	(0.34)	23.26	(0.33)	22.22	(0.32)	22.48	(0.32)
	CP-GLM	24.66	(0.36)	24.42	(0.35)	23.10	(0.34)	23.50	(0.35)
	ℓ_1-GLM	24.20	(0.32)	23.09	(0.35)	21.68	(0.33)	21.77	(0.33)
	ℓ_1-FDA	26.11	(0.38)	23.12	(0.35)	22.17	(0.33)	22.21	(0.33)
Distance Correlation	CATCH	23.66	(0.34)	23.31	(0.34)	22.13	(0.31)	22.35	(0.31)
	CP-GLM	24.66	(0.36)	25.02	(0.35)	22.61	(0.34)	22.15	(0.34)
	ℓ_1-GLM	24.20	(0.32)	23.07	(0.33)	21.41	(0.33)	21.78	(0.32)
	ℓ_1-FDA	26.11	(0.38)	23.23	(0.34)	21.90	(0.32)	22.31	(0.32)

the smoothness among electrodes. Another interesting fact is that, even without smoothness along the columns, STS continues to give comparable error rates to marginal screening. This is likely because we still have smoothness along the time domain, and, as noted in Section 5.1, STS is capable of exploiting partial smoothness.

Moreover, the average computation time to perform screening is reported in Table 6, which confirms that STS is as computationally efficient as the corresponding marginal screening methods.

6. Conclusion

In this article, we propose STS, a general screening framework for tensors. STS integrates the traditional marginal screening methods with the tensor structural information. With a wide selection of statistical utilities in screening, STS is not limited to any model setting or data type of the responses and predictors. We establish the SURE screening property for the procedure and give three examples. Moreover, we examine the performance by comparing the classification error and regression accuracy on screened data. STS gives better results than directly applying traditional screening methods on the vectorized tensor in both simulation and real data study. In practice, researchers can combine STS with other suitable screening utilities to improve their performance on tensor data. But an exhaustive study along this direction is apparently out of the scope of this article.

TABLE 5
The means and standard errors of binary classification error rates on EEG dataset after shuffling electrodes. We only report the results for STS because the results without screening and those with marginal screening are the same as in Table 4 and omitted.

Screening Method		$c = \omega/2$		$c = \omega$	
		Mean	S.E.	Mean	S.E.
t-statistic	CATCH	22.87	(0.32)	22.74	(0.31)
	CP-GLM	23.43	(0.32)	23.36	(0.33)
	ℓ_1-GLM	23.01	(0.32)	22.82	(0.32)
	ℓ_1-FDA	23.61	(0.33)	24.33	(0.33)
Logistic Regression	CATCH	22.34	(0.31)	22.42	(0.32)
	CP-GLM	23.67	(0.34)	23.61	(0.34)
	ℓ_1-GLM	22.81	(0.33)	22.46	(0.33)
	ℓ_1-FDA	22.86	(0.32)	23.35	(0.32)
Distance Correlation	CATCH	22.50	(0.31)	22.39	(0.31)
	CP-GLM	23.60	(0.32)	23.72	(0.33)
	ℓ_1-GLM	22.71	(0.31)	22.49	(0.32)
	ℓ_1-FDA	23.08	(0.32)	23.73	(0.32)

TABLE 6
Average computation time for 100 replicates. The STS has similar computation cost as marginal screening.

Time (s)	Marginal	$c = \omega/2$	$c = \omega$
t-statistic	2.18	2.19	2.18
Logistic Regression	19.16	19.15	19.16
Distance Correlation	9.71	9.71	9.71

If we are fitting a specific kind of tensor model, it is also possible to modify the screening utilities to further take advantage of the tensor structure (with or without the smoothness assumption). For example, under the TDA model, the variance $\sigma_{k\mathcal{J}}$ can be estimated much more accurately [33, 42, 40] than the sample estimate. The improved estimation could benefit screening. We tested this idea on our simulations models, and the results are comparable to those of STS, so we do not report them. However, it is still likely that screening utilities calculated utilizing the tensor structure will be helpful in other settings or under other models. It is a topic worth exploring in future research.

We focus on the problem where the predictor is a tensor, but the response is a scalar. There are considerable interests in the literature where the response is a tensor as well [20, 31, 29, 15, 26, 30, 1, 27, e.g.]. These papers generally study fitting tensor-on-tensor regression models. It will be interesting to develop screening methods under such models as well, but such developments are out of the scope of this paper for two reasons. On one hand, it is unclear how to

generalize the smoothness assumption to tensor-on-tensor problems. There are several possibilities that could be explored, such as smoothness in the response alone, in the predictor alone, or both. On the other hand, it is critical for screening methods to enjoy the SURE screening property under ultra-high dimensions, but the aforementioned tensor-on-tensor works either have few results on the statistical properties, or only have results in low-dimensional problems. Hence, full developments on the theory could be challenging for screening methods in the same context. A relevant screening method is proposed in [22], where the response is a matrix (two-way tensor), and the predictor is a vector. Our problem is different in that our predictor, instead of the response, is a tensor of arbitrary order. Nevertheless, it will be intriguing to investigate in the future whether some of their results can assist in tensor-on-tensor screening.

Appendix A: Proofs

A.1. Proof of Theorem 4.1

Proof of Theorem 4.1. We first give proof of the first conclusion. By definition, for $0 < \epsilon < \epsilon_0$, we have

$$\begin{aligned}
&\operatorname{pr}\left\{\left|\phi_{n\mathcal{J}}^{\text{Smooth}} - \phi_{\mathcal{J}}^{\text{Smooth}}\right| > (1+c)\epsilon\right\} \\
=&\operatorname{pr}\left\{\left|\left|\phi_{n\mathcal{J}}\right| - \left|\phi_{\mathcal{J}}\right| + \frac{c}{|\Omega_{\mathcal{J}}|}\sum_{\mathcal{I}\in\Omega_{\mathcal{J}}}\left(\left|\phi_{n\mathcal{I}}\right| - \left|\phi_{\mathcal{I}}\right|\right)\right| > (1+c)\epsilon\right\} \\
\leq&\operatorname{pr}\left(\left|\phi_{n\mathcal{J}} - \phi_{\mathcal{J}}\right| > \epsilon\right) + \operatorname{pr}\left(\frac{c}{|\Omega_{\mathcal{J}}|}\sum_{\mathcal{I}\in\Omega_{\mathcal{J}}}\left|\phi_{n\mathcal{I}} - \phi_{\mathcal{I}}\right| > c\epsilon\right) \\
\leq&\operatorname{pr}\left(\left|\phi_{n\mathcal{J}} - \phi_{\mathcal{J}}\right| > \epsilon\right) + \sum_{\mathcal{I}\in\Omega_{\mathcal{J}}}\operatorname{pr}\left(\left|\phi_{n\mathcal{I}} - \phi_{\mathcal{I}}\right| > \epsilon\right) \\
\leq&(1+|\Omega_{\mathcal{J}}|)\operatorname{pr}\left(\left|\phi_{n\mathcal{J}} - \phi_{\mathcal{J}}\right| > \epsilon\right) \\
\leq&(1+\omega)\zeta_n(\epsilon).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\operatorname{pr}\left\{\max_{\mathcal{J}}\left|\phi_{n\mathcal{J}}^{\text{Smooth}} - \phi_{\mathcal{J}}^{\text{Smooth}}\right| > (1+c)\epsilon\right\} \\
\leq&\sum_{\mathcal{J}}\operatorname{pr}\left\{\left|\phi_{n\mathcal{J}}^{\text{Smooth}} - \phi_{\mathcal{J}}^{\text{Smooth}}\right| > (1+c)\epsilon\right\} \\
\leq&\Big(\prod_{r=1}^{R} p_r\Big)(1+\omega)\zeta_n(\epsilon). \qquad \text{(A.1)}
\end{aligned}$$

The proof of the first part is finished.

Next, we prove the second part. Let $\epsilon_1 = \min\{\epsilon_0, \delta_{\mathcal{S}_1}^{\text{Smooth}}/(2(1+c))\}$, using (A.1), we have

$$\begin{aligned}
&\text{pr}\left(\max_{\mathcal{J}} |\phi_{n\mathcal{J}}^{\text{Smooth}} - \phi_{\mathcal{J}}^{\text{Smooth}}| > \delta_{\mathcal{S}_1}^{\text{Smooth}}/2\right) \\
&\leq \sum_{\mathcal{J}} \text{pr}\left(|\phi_{n\mathcal{J}}^{\text{Smooth}} - \phi_{\mathcal{J}}^{\text{Smooth}}| > \delta_{\mathcal{S}_1}^{\text{Smooth}}/2\right) \\
&\leq \sum_{\mathcal{J}} \text{pr}\left(|\phi_{n\mathcal{J}}^{\text{Smooth}} - \phi_{\mathcal{J}}^{\text{Smooth}}| > (1+c)\epsilon_1\right) \\
&\leq (\prod_{r=1}^{R} p_r)(1+\omega)\zeta_n(\epsilon_1).
\end{aligned}$$

Then $\text{pr}\left\{\mathcal{D} \subseteq \widehat{\mathcal{S}}^{\text{Smooth}}(d_n)\right\} \geq 1 - (\prod_{r=1}^{R} p_r)(1+\omega)\zeta_n(\epsilon_1)$. □

A.2. Proof of Lemma 4.1

Proof of Lemma 4.1. Under the setting in Assumption (A1), we let $\mathcal{S} = \cup_{j \leq J_0} \mathcal{B}_j$. The set $\mathcal{S}$ satisfies Condition (V1) with $\delta_{\mathcal{S}} = \min_{j \leq J_0} v_j > 0$. Denote $v_{\mathcal{S},\max} = \max_{j \leq J_0} v_j$, $v_{\mathcal{S},\min} = \min_{j \leq J_0} v_j$ and $v_{\mathcal{S}^c} = 0$. We consider the set $\mathcal{S}_1 = \mathcal{S} \cup \{\cup_{\mathcal{J} \in \mathcal{S}} \Omega_{\mathcal{J}}\}$. If $\mathcal{J} \in \mathcal{S}_1$,

$$\phi_{\mathcal{J}}^{\text{Smooth}} > \min\{v_{\mathcal{S},\min} + c v_{\mathcal{S}^c}, v_{\mathcal{S}^c} + \frac{c}{|\Omega_{\mathcal{J}}|} v_{\mathcal{S},\min} + \frac{c(|\Omega_{\mathcal{J}}| - 1)}{|\Omega_{\mathcal{J}}|} v_{\mathcal{S}^c}\}.$$

If $\mathcal{J} \in \mathcal{S}_1^c$,

$$\phi_{\mathcal{J}}^{\text{Smooth}} = (1+c) v_{\mathcal{S}^c}.$$

Thus,

$$\delta_{\mathcal{S}_1}^{\text{Smooth}} = \min_{\mathcal{J} \in \mathcal{S}_1} \{\phi_{\mathcal{J}}^{\text{Smooth}}\} - \max_{\mathcal{J} \in \mathcal{S}_1^c} \{\phi_{\mathcal{J}}^{\text{Smooth}}\} > \min\{1, \frac{c}{\omega}\} \delta_{\mathcal{S}}.$$

Next, we derive the upper bound for $|\mathcal{S}_1|$. Suppose that $\mathcal{B}_j \in \mathbb{R}^{d_1 \times \cdots \times d_R}$, then $|\mathcal{B}_j| = d_1 \times \cdots \times d_R$. We use $\mathcal{B}_j'$ to denote the set containing all elements in $\mathcal{B}_j$ and their neighbors. By definition, we have that $|\mathcal{B}_j'| = (d_1 + 2) \times \cdots \times (d_R + 2)$. For fixed $|\mathcal{B}_j|$, when $d_k = |\mathcal{B}_j|$ for some k and $d_i = 1$ for all $i \neq k$, $|\mathcal{B}_j'|$ reaches its maximum at $3^{R-1}(|\mathcal{B}_j| + 2)$. Therefore, we have $|\mathcal{S}_1| \leq \sum_{j \leq J_0} |\mathcal{B}_j'| \leq 3^{R-1}(|\mathcal{S}| + 2J_0)$. □

A.3. Proof of Corollary 4.1

The proof of Corollary 4.1 relies on the following proposition. We first prove Proposition A.1 and then prove Corollary 4.1.

Proposition A.1. *Under Condition (T3), we have that* $\text{pr}(\max_{\mathcal{J}} |\phi^t_{n\mathcal{J}} - \phi^t_{\mathcal{J}}| > \epsilon) \leq (\prod_{r=1}^{R} p_r)\zeta^t_n(\epsilon)$*, where*

$$\zeta^t_n(\epsilon) = \gamma_1 \exp(-\gamma_2 n\epsilon^2) + \gamma_3 \exp(-\gamma_4 n),$$

for some positive constants γ_1*,* γ_2*,* γ_3 *and* γ_4*.*

By Proposition A.1, Condition (T3) implies Condition (T2). Condition (T3) is a widely used assumption in high-dimensional statistics. Hence, this demonstrates that Condition (T2) is indeed very mild.

The proof of Proposition A.1 is straightforward, but we include it for completeness. To prove Proposition A.1, we start with some propositions that are used in our proofs and they are extracted from [52].

Proposition A.2 (Hoeffding bound, [52], cf. Proposition 2.5)**.** *Suppose that the variables* $X_i, i = 1, \ldots, n$*, are independent, and* X_i *has mean* μ_i *and sub-Gaussian parameter* σ_i*. Then for all* $t \geq 0$*, we have*

$$\text{pr}\left\{\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right\} \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Proposition A.3 ([52], cf. Proposition 2.9)**.** *Suppose that the random variable* X *with mean* $\mu = \text{E}(X)$ *is sub-exponential with parameters* (v, α)*. Then*

$$\text{pr}(X - \mu \geq t) \leq \begin{cases} e^{-\frac{t^2}{2v^2}} & \text{if } 0 \leq t \leq \frac{v^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}} & \text{for } t > \frac{v^2}{\alpha}. \end{cases}$$

Proposition A.4 ([52], cf. equation (2.18))**.** *Suppose that* $\{X_k\}_{k=1}^n$ *is an independent sequence of random variables, such that* X_k *has mean* μ_k*, and is sub-exponential with parameters* (v_k, α_k)*, then the variable* $\sum_{k=1}^n (X_k - \mu_k)$ *is sub-exponential with the parameters* (v_*, α_*)*, where*

$$\alpha_* = \max_{k=1,\ldots,n} \alpha_k \quad \text{and} \quad v_* = \left(\sum_{k=1}^{n} v_k^2\right)^{1/2}.$$

And we have the tail bound

$$\text{pr}\left\{\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \geq t\right\} \leq \begin{cases} e^{-\frac{nt^2}{2(v_*^2/n)}} & \text{for } 0 \leq t \leq \frac{v_*^2}{n\alpha_*}, \\ e^{-\frac{nt}{2\alpha_*}} & \text{for } t > \frac{v_*^2}{n\alpha_*}. \end{cases}$$

Proposition A.5. *Let* $X \sim \text{subG}\left(\sigma^2\right)$*. Then the random variable* $Z = X^2 - \text{E}\left(X^2\right)$ *is sub-exponential:* $Z \sim \text{subE}\left(\psi_1\sigma^2, \psi_2\sigma^2\right)$ *with positive constants* ψ_1 *and* ψ_2*.*

The screening utility is defined as

$$\phi^t_{n\mathcal{J}} = \frac{|\overline{X}_{1\mathcal{J}} - \overline{X}_{2\mathcal{J}}|}{n^{1/2}\left(\frac{\widehat{\sigma}^2_{1\mathcal{J}}}{n_1} + \frac{\widehat{\sigma}^2_{2\mathcal{J}}}{n_2}\right)^{1/2}} = \frac{|\overline{X}_{1\mathcal{J}} - \overline{X}_{2\mathcal{J}}|}{\left(\frac{\widehat{\sigma}^2_{1\mathcal{J}}}{n_1/n} + \frac{\widehat{\sigma}^2_{2\mathcal{J}}}{n_2/n}\right)^{1/2}},$$

where $\widehat{\sigma}^2_{k\mathcal{J}}$ is the sample variance of variable $\mathcal{J}$ within class k. Recall the screening utility and its population counterpart defined in (3.5) & (3.6). We further define an intermediate quantity

$$\phi^{t\ (1)}_{\mathcal{J}} = \frac{|\mu_{1\mathcal{J}} - \mu_{2\mathcal{J}}|}{\left(\frac{\sigma^2_{1\mathcal{J}}}{n_1/n} + \frac{\sigma^2_{2\mathcal{J}}}{n_2/n}\right)^{1/2}}.$$

To make the proof easier to read, we introduce the following shorthand notations. For a given $\mathcal{J}$, let

$$a_{n\mathcal{J}} = \overline{X}_{1\mathcal{J}} - \overline{X}_{2\mathcal{J}}, \qquad a_{\mathcal{J}} = \mu_{1\mathcal{J}} - \mu_{2\mathcal{J}},$$
$$b_{n\mathcal{J}} = \left(\frac{\widehat{\sigma}^2_{1\mathcal{J}}}{n_1/n} + \frac{\widehat{\sigma}^2_{2\mathcal{J}}}{n_2/n}\right)^{1/2}, \qquad b_{\mathcal{J}} = \left(\frac{\sigma^2_{1\mathcal{J}}}{n_1/n} + \frac{\sigma^2_{2\mathcal{J}}}{n_2/n}\right)^{1/2}.$$

For Condition (T3), we use positive constants a, $\sigma^2_{1,\min}$, $\sigma^2_{1,\max}$, $\sigma^2_{2,\min}$ and $\sigma^2_{2,\max}$ to define the bounds. Then we have

$$|a_{\mathcal{J}}| \le a, \quad \sigma^2_{1,\min} \le \sigma^2_{1\mathcal{J}} \le \sigma^2_{1,\max}, \quad \sigma^2_{2,\min} \le \sigma^2_{2\mathcal{J}} \le \sigma^2_{2,\max}.$$

In the proof, we use $\gamma_i, i = 1, 2, \ldots$ to denote positive constants. They can have different value each time they appear.

Proof of Proposition A.1. Since we have

$$\begin{aligned}\operatorname{pr}\left(\left|\phi^t_{n\mathcal{J}} - \phi^t_{\mathcal{J}}\right| > \epsilon\right) &\le \operatorname{pr}\left(\left|\phi^t_{n\mathcal{J}} - \phi^{t\ (1)}_{\mathcal{J}}\right| > \frac{\epsilon}{2}\right) + \operatorname{pr}\left(\left|\phi^{t\ (1)}_{\mathcal{J}} - \phi^t_{\mathcal{J}}\right| > \frac{\epsilon}{2}\right) \\ &= L_1 + L_2,\end{aligned}$$

we will bound the probabilities L_1 and L_2 separately.

For L_1,

$$|\phi^t_{n\mathcal{J}} - \phi^{t\ (1)}_{\mathcal{J}}| \le \left|\frac{a_{n\mathcal{J}}}{b_{n\mathcal{J}}} - \frac{a_{\mathcal{J}}}{b_{\mathcal{J}}}\right| \le \frac{|a_{n\mathcal{J}} - a_{\mathcal{J}}|}{b_{n\mathcal{J}}} + \frac{|a_{\mathcal{J}}|}{b_{n\mathcal{J}} b_{\mathcal{J}}}|b_{n\mathcal{J}} - b_{\mathcal{J}}|. \tag{A.2}$$

First, we give an upper bound to $|a_{n\mathcal{J}} - a_{\mathcal{J}}|$. Note that

$$\begin{aligned}&\operatorname{pr}\left(|a_{n\mathcal{J}} - a_{\mathcal{J}}| > \epsilon \mid Y\right) \\ &= \operatorname{pr}\left(\left|\overline{X}_{1\mathcal{J}} - \overline{X}_{2\mathcal{J}} - (\mu_{1\mathcal{J}} - \mu_{2\mathcal{J}})\right| > \epsilon \mid Y\right) \\ &\le \operatorname{pr}\left(\left|\overline{X}_{1\mathcal{J}} - \mu_{1\mathcal{J}}\right| > \frac{\epsilon}{2} \mid Y\right) + \operatorname{pr}\left(\left|\overline{X}_{2\mathcal{J}} - \mu_{2\mathcal{J}}\right| > \frac{\epsilon}{2} \mid Y\right) \qquad \text{(A.3)} \\ &\le 2\exp\left(-\frac{n_1\epsilon^2}{8\sigma^2_{1\mathcal{J}}}\right) + 2\exp\left(-\frac{n_2\epsilon^2}{8\sigma^2_{2\mathcal{J}}}\right), \qquad \text{(A.4)}\end{aligned}$$

where we use Proposition A.2 to get (A.4) from (A.3).

As n_1 and n_2 are sums of n independent and identically distributed Bernoulli random variables, by Hoeffding's inequality, we have

$$\operatorname{pr}\left(|n_1 - n\pi_1| > \frac{n\pi_1}{2}\right) < 2\exp\left(-\frac{n\pi_1^2}{2}\right),$$

$$\mathrm{pr}\left(|n_2 - n\pi_2| > \frac{n\pi_2}{2}\right) < 2\exp\left(-\frac{n\pi_2^2}{2}\right).$$

Let $A = \{n_1, n_2 : \frac{n\pi_1}{2} < n_1 < \frac{3n\pi_1}{2}, \frac{n\pi_2}{2} < n_2 < \frac{3n\pi_2}{2}\}$, we have

$$\begin{aligned}
&\mathrm{pr}\left(|a_{n\mathcal{J}} - a_{\mathcal{J}}| > \epsilon\right) \\
&= \mathrm{E}\left[\mathrm{pr}\left\{\left|\overline{X}_{1\mathcal{J}} - \overline{X}_{2\mathcal{J}} - (\mu_{1\mathcal{J}} - \mu_{2\mathcal{J}})\right| > \epsilon \mid Y\right\}\right] \\
&\le \mathrm{E}\left[\left\{2\exp\left(-\frac{n_1\epsilon^2}{8\sigma_{1\mathcal{J}}^2}\right) + 2\exp\left(-\frac{n_2\epsilon^2}{8\sigma_{2\mathcal{J}}^2}\right)\right\} \cdot \mathbf{1}_A\right] + 2\mathrm{E}(\mathbf{1}_{A^c}) \\
&\le 2\exp\left(-\frac{n\pi_1\epsilon^2}{16\sigma_{1\mathcal{J}}^2}\right) + 2\exp\left(-\frac{n\pi_2\epsilon^2}{16\sigma_{2\mathcal{J}}^2}\right) \\
&+ 2\,\mathrm{pr}\left(|n_1 - n\pi_1| > \frac{n\pi_1}{2}\right) + 2\,\mathrm{pr}\left(|n_2 - n\pi_2| > \frac{n\pi_2}{2}\right) \\
&\le \gamma_1\exp\left(-\gamma_2 n\epsilon^2\right) + \gamma_3\exp\left(-\gamma_4 n\right). \qquad \text{(A.5)}
\end{aligned}$$

Next, we give an upper bound to $|b_{n\mathcal{J}} - b_{\mathcal{J}}|$,

$$\begin{aligned}
&\mathrm{pr}\left(|b_{n\mathcal{J}} - b_{\mathcal{J}}| > \epsilon \mid Y\right) \\
&\le \mathrm{pr}\left(\frac{\left|b_{n\mathcal{J}}^2 - b_{\mathcal{J}}^2\right|}{b_{n\mathcal{J}} + b_{\mathcal{J}}} > \epsilon \mid Y\right) \\
&\le \mathrm{pr}\left(\frac{\left|b_{n\mathcal{J}}^2 - b_{\mathcal{J}}^2\right|}{b_{\mathcal{J}}} > \epsilon \mid Y\right) \\
&\le \mathrm{pr}\left(\left|\frac{\widehat{\sigma}_{1\mathcal{J}}^2 - \sigma_{1\mathcal{J}}^2}{n_1/n}\right| > \frac{b_{\mathcal{J}}\epsilon}{2} \mid Y\right) + \mathrm{pr}\left(\left|\frac{\widehat{\sigma}_{2\mathcal{J}}^2 - \sigma_{2\mathcal{J}}^2}{n_2/n}\right| > \frac{b_{\mathcal{J}}\epsilon}{2} \mid Y\right).
\end{aligned}$$

By definition, we have

$$\begin{aligned}
&\widehat{\sigma}_{1\mathcal{J}}^2 - \sigma_{1\mathcal{J}}^2 \\
&= \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1\mathcal{J}}^i - \overline{X}_{1\mathcal{J}})^2 - \sigma_{1\mathcal{J}}^2 \\
&= \frac{1}{n_1 - 1}\left\{\sum_{i=1}^{n_1}(X_{1\mathcal{J}}^i - \mu_{1\mathcal{J}})^2 - n_1(\overline{X}_{1\mathcal{J}} - \mu_{1\mathcal{J}})^2\right\} - \sigma_{1\mathcal{J}}^2 \\
&= \frac{n_1}{n_1 - 1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\left(X_{1\mathcal{J}}^i - \mu_{1\mathcal{J}}\right)^2 - \sigma_{1\mathcal{J}}^2\right\} - \left\{\left(\overline{X}_{1\mathcal{J}} - \mu_{1\mathcal{J}}\right)^2 - \frac{\sigma_{1\mathcal{J}}^2}{n_1}\right\}\right] \\
&= \frac{n_1}{n_1 - 1}\left(I_1 - I_2\right).
\end{aligned}$$

Hence,

$$\mathrm{pr}\left(\left|\frac{\widehat{\sigma}_{1\mathcal{J}}^2 - \sigma_{1\mathcal{J}}^2}{n_1/n}\right| > \frac{b_{\mathcal{J}}\epsilon}{2} \mid Y\right)$$

$$
\begin{aligned}
&= \mathrm{pr}\left(|I_1 - I_2| > \frac{n_1 - 1}{n}\frac{b_{\mathcal{J}}\epsilon}{2} \mid Y\right) \\
&\le \mathrm{pr}\left(|I_1| > \frac{\widehat{\pi}_1 b_{\mathcal{J}}\epsilon}{8} \mid Y\right) + \mathrm{pr}\left(|I_2| > \frac{\widehat{\pi}_1 b_{\mathcal{J}}\epsilon}{8} \mid Y\right).
\end{aligned}
$$

Under Condition (T3), $X^i_{1\mathcal{J}} - \mu_{1\mathcal{J}} \sim \mathrm{subG}(\sigma^2_{1\mathcal{J}})$. In addition, $\overline{X}_{1\mathcal{J}} - \mu_{1\mathcal{J}} \sim \mathrm{subG}(\frac{\sigma^2_{1\mathcal{J}}}{n_1})$. Using Proposition A.5, we can show that $\left(X^i_{1\mathcal{J}} - \mu_{1\mathcal{J}}\right)^2 - \sigma^2_{1\mathcal{J}} \sim \mathrm{subE}(\psi_1\sigma^2_{1\mathcal{J}}, \psi_2\sigma^2_{1\mathcal{J}})$ and $\left(\overline{X}_{1\mathcal{J}} - \mu_{1\mathcal{J}}\right)^2 - \frac{\sigma^2_{1\mathcal{J}}}{n_1} \sim \mathrm{subE}(\psi_3\frac{\sigma^2_{1\mathcal{J}}}{n_1}, \psi_4\frac{\sigma^2_{1\mathcal{J}}}{n_1})$, where $\psi_1, \psi_2, \psi_3, \psi_4$ are positive constants. Then by Propositions A.3 and A.4, we have

$$
\begin{aligned}
\mathrm{pr}\left(|I_1| > \frac{\widehat{\pi}_1 b_{\mathcal{J}}\epsilon}{8} \mid Y\right) &\le 2\exp\left(-\min\{g_{1\mathcal{J}} n\epsilon^2, g_{2\mathcal{J}} n\epsilon\}\right), \\
\mathrm{pr}\left(|I_2| > \frac{\widehat{\pi}_1 b_{\mathcal{J}}\epsilon}{8} \mid Y\right) &\le 2\exp\left(-\min\{h_{1\mathcal{J}} n^2\epsilon^2, h_{2\mathcal{J}} n\epsilon\}\right),
\end{aligned}
$$

where $g_{1\mathcal{J}}, g_{2\mathcal{J}}, h_{1\mathcal{J}}, h_{2\mathcal{J}}$ are functions of $\sigma^2_{1\mathcal{J}}, \sigma^2_{2\mathcal{J}}, \widehat{\pi}_1, \widehat{\pi}_2$ that are not related to n or p. Recall that $A = \{n_1, n_2 : \frac{n\pi_1}{2} < n_1 < \frac{3n\pi_1}{2}, \frac{n\pi_2}{2} < n_2 < \frac{3n\pi_2}{2}\}$. When A holds and under Condition (T3), the four functions are bounded as well. Therefore, for small ϵ, we have

$$
\begin{aligned}
&\mathrm{pr}\left(\left|\frac{\widehat{\sigma}^2_{1\mathcal{J}} - \sigma^2_{1\mathcal{J}}}{n_1/n}\right| > \frac{b_{\mathcal{J}}\epsilon}{2} \mid A\right) \\
&\le 2\exp\left(-\min\{\gamma_1 n\epsilon^2, \gamma_2 n\epsilon\}\right) + 2\exp\left(-\min\{\gamma_3 n^2\epsilon^2, \gamma_4 n\epsilon\}\right) \\
&\le \gamma_1\exp\left(-\gamma_2 n\epsilon^2\right).
\end{aligned}
$$

We can derive an upper bound for $|\widehat{\sigma}^2_{2\mathcal{J}} - \sigma^2_{2\mathcal{J}}|$ in the same way. By reorganizing the constants, we finally get

$$
\begin{aligned}
&\mathrm{pr}\left(|b_{n\mathcal{J}} - b_{\mathcal{J}}| > \epsilon\right) \\
&= \mathrm{E}\{\mathrm{pr}\left(|b_{n\mathcal{J}} - b_{\mathcal{J}}| > \epsilon \mid Y\right)\} \\
&\le \mathrm{pr}\left(\left|\frac{\widehat{\sigma}^2_{1\mathcal{J}} - \sigma^2_{1\mathcal{J}}}{n_1/n}\right| > \frac{b_{\mathcal{J}}\epsilon}{2} \mid A\right) + \mathrm{pr}\left(\left|\frac{\widehat{\sigma}^2_{2\mathcal{J}} - \sigma^2_{2\mathcal{J}}}{n_2/n}\right| > \frac{b_{\mathcal{J}}\epsilon}{2} \mid A\right) + 2E\{\mathbf{1}_{A^c}\} \\
&\le \gamma_1\exp\left(-\gamma_2 n\epsilon^2\right) + \gamma_3\exp\left(-\gamma_4 n\right). \qquad \text{(A.6)}
\end{aligned}
$$

Let $|b_{n\mathcal{J}} - b_{\mathcal{J}}| < \epsilon < \frac{b_{\mathcal{J}}}{2}$, then $b_{n\mathcal{J}} > \frac{b_{\mathcal{J}}}{2}$. When event A holds, we can find a constant u such that $b_{\mathcal{J}} \ge u$ for all $\mathcal{J}$. Thus, under the conditions $|a_{n\mathcal{J}} - a_{\mathcal{J}}| < \epsilon$ and $|b_{n\mathcal{J}} - b_{\mathcal{J}}| < \epsilon < \frac{b_{\mathcal{J}}}{2}$, (A.2) becomes

$$
|\phi^t_{n\mathcal{J}} - {\phi^t_{\mathcal{J}}}^{(1)}| < \frac{2}{b_{\mathcal{J}}}\epsilon + \frac{2|a_{\mathcal{J}}|}{b^2_{\mathcal{J}}}\epsilon < \left(\frac{2}{u} + \frac{2|a|}{u^2}\right)\epsilon = \eta_1\epsilon. \qquad \text{(A.7)}
$$

Define $b_{\min} = \min_{\mathcal{J}} b_{\mathcal{J}}$. Combining the results (A.5), (A.6) and (A.7), we have for any $0 < \epsilon < \frac{b_{\min}}{2}$ and fixed $\mathcal{J}$,

$$
\mathrm{pr}\left(|\phi^t_{n\mathcal{J}} - {\phi^t_{\mathcal{J}}}^{(1)}| < \eta_1\epsilon\right)
$$

$$\begin{aligned}&\geq \mathrm{pr}\left(\{|b_{n\mathcal{J}} - b_{\mathcal{J}}| < \epsilon\} \cap \{|a_{n\mathcal{J}} - a_{\mathcal{J}}| < \epsilon\} \cap A\right)\\ &\geq 1 - \mathrm{pr}\left(|b_{n\mathcal{J}} - b_{\mathcal{J}}| > \epsilon\right) - \mathrm{pr}\left(|a_{n\mathcal{J}} - a_{\mathcal{J}}| > \epsilon\right) - \mathrm{pr}(A^c)\\ &\geq 1 - \gamma_1 \exp(-\gamma_2 n\epsilon^2) - \gamma_3 \exp(-\gamma_4 n).\end{aligned}$$

Let $\frac{\nu}{2} = \eta_1\epsilon$. By reorganizing the constants, we have that for any $0 < \nu < \min\{\eta_1 b_{\min}, 2\eta_1\}$,

$$\mathrm{pr}\left(|\phi^t_{n\mathcal{J}} - \phi^{t\,(1)}_{\mathcal{J}}| > \frac{\nu}{2}\right) \leq \gamma_1 \exp(-\gamma_2 n\nu^2) + \gamma_3 \exp(-\gamma_4 n). \tag{A.8}$$

Next, we consider L_2. Let $b_{0\mathcal{J}} = \left(\frac{\sigma^2_{1\mathcal{J}}}{\pi_1} + \frac{\sigma^2_{2\mathcal{J}}}{\pi_2}\right)^{1/2} \geq \left(\frac{\sigma^2_{1,\min}}{\pi_1} + \frac{\sigma^2_{2,\min}}{\pi_2}\right)^{1/2} = b_{0,\min}$, then

$$\phi^{t\,(1)}_{\mathcal{J}} = \frac{|a_{\mathcal{J}}|}{b_{\mathcal{J}}}, \quad \phi^t_{\mathcal{J}} = \frac{|a_{\mathcal{J}}|}{b_{0\mathcal{J}}}.$$

Let $|\widehat{\pi}_1 - \pi_1| < \epsilon < \frac{\pi_1}{2}$, $|\widehat{\pi}_2 - \pi_2| < \epsilon < \frac{\pi_2}{2}$, then we have $\widehat{\pi}_1 > \frac{\pi_1}{2}$ and $\sqrt{\frac{2}{3}} b_{0\mathcal{J}} < b_{\mathcal{J}} < \sqrt{2} b_{0\mathcal{J}}$. Under these conditions, we can get

$$|\phi^{t\,(1)}_{\mathcal{J}} - \phi^t_{\mathcal{J}}| \leq \left|\frac{a_{\mathcal{J}}}{b_{\mathcal{J}}} - \frac{a_{\mathcal{J}}}{b_{0\mathcal{J}}}\right| = |a_{\mathcal{J}}| \frac{|b_{\mathcal{J}} - b_{0\mathcal{J}}|}{b_{\mathcal{J}} b_{0\mathcal{J}}} < \frac{|a_{\mathcal{J}}|}{\sqrt{\frac{2}{3}} b^2_{0\mathcal{J}}} |b_{\mathcal{J}} - b_{0\mathcal{J}}|. \tag{A.9}$$

Furthermore, we have

$$|b_{\mathcal{J}} - b_{0\mathcal{J}}| \leq \frac{\left|b^2_{\mathcal{J}} - b^2_{0\mathcal{J}}\right|}{b_{0\mathcal{J}}} \leq \frac{1}{b_{0\mathcal{J}}} \left(\sigma^2_{1\mathcal{J}} \left|\frac{1}{\widehat{\pi}_1} - \frac{1}{\pi_1}\right| + \sigma^2_{2\mathcal{J}} \left|\frac{1}{\widehat{\pi}_2} - \frac{1}{\pi_2}\right|\right), \tag{A.10}$$

and

$$\left|\frac{1}{\widehat{\pi}_1} - \frac{1}{\pi_1}\right| = \frac{|\widehat{\pi}_1 - \pi_1|}{\pi_1 \widehat{\pi}_1} < \frac{2}{\pi^2_1}\epsilon, \quad \left|\frac{1}{\widehat{\pi}_2} - \frac{1}{\pi_2}\right| = \frac{|\widehat{\pi}_2 - \pi_2|}{\pi_2 \widehat{\pi}_2} < \frac{2}{\pi^2_2}\epsilon. \tag{A.11}$$

Combining the results in (A.9), (A.10) and (A.11), we obtain

$$\begin{aligned}|\phi^{t\,(1)}_{\mathcal{J}} - \phi^t_{\mathcal{J}}| &< \frac{|a_{\mathcal{J}}|}{\sqrt{\frac{2}{3}} b^2_{0\mathcal{J}}} \frac{2}{b_{0\mathcal{J}}} \left(\frac{\sigma^2_{1\mathcal{J}}}{\pi^2_1} + \frac{\sigma^2_{2\mathcal{J}}}{\pi^2_2}\right)\epsilon\\ &\leq \frac{2|a|}{\sqrt{\frac{2}{3}} b^3_{0,\min}} \left(\frac{\sigma^2_{1,\max}}{\pi^2_1} + \frac{\sigma^2_{2,\max}}{\pi^2_2}\right)\epsilon = \eta_2\epsilon.\end{aligned}$$

Thus, for any given $\epsilon < \min\{\frac{\pi_1}{2}, \frac{\pi_2}{2}\}$,

$$\begin{aligned}\mathrm{pr}\left(|\phi^t_{n\mathcal{J}} - \phi^{t\,(1)}_{\mathcal{J}}| < \eta_2\epsilon\right) &\geq \mathrm{pr}\left(\{|\widehat{\pi}_1 - \pi_1| < \epsilon\} \cap \{|\widehat{\pi}_2 - \pi_2| < \epsilon\}\right)\\ &\geq 1 - \mathrm{pr}\left(|\widehat{\pi}_1 - \pi_1| > \epsilon\right) - \mathrm{pr}\left(|\widehat{\pi}_2 - \pi_2| > \epsilon\right)\\ &\geq 1 - 4\exp\left(-2n\epsilon^2\right).\end{aligned} \tag{A.12}$$

Let $\frac{\nu}{2} = \eta_2\epsilon$, then (A.12) is equivalent to

$$\text{pr}\left(|\phi^t_{n\mathcal{J}} - \phi^{t\ (1)}_{\mathcal{J}}| > \frac{\nu}{2}\right) \le \gamma_5 \exp\left(-\gamma_6 n\nu^2\right), \text{ for any } 0 < \nu < \min\{\eta_2\pi_1, \eta_2\pi_2\}. \tag{A.13}$$

Finally, we combine the results (A.8) and (A.13). Then for any ϵ such that $0 < \epsilon < \epsilon_0 = \min\{\eta_1 b_{\min}, 2\eta_1, \eta_2\pi_1, \eta_2\pi_2\}$, we have

$$\begin{aligned}
\text{pr}\left(|\phi^t_{n\mathcal{J}} - \phi^t_{\mathcal{J}}| > \epsilon\right) &\le \text{pr}\left(\left|\phi^t_{n\mathcal{J}} - \phi^{t\ (1)}_{\mathcal{J}}\right| > \epsilon/2\right) + \text{pr}\left(\left|\phi^{t\ (1)}_{\mathcal{J}} - \phi^t_{\mathcal{J}}\right| > \epsilon/2\right) \\
&= \gamma_1 \exp(-\gamma_2 n\epsilon^2) + \gamma_3 \exp(-\gamma_4 n) + \gamma_5 \exp\left(-\gamma_6 n\epsilon^2\right) \\
&\le \gamma_1 \exp(-\gamma_2 n\epsilon^2) + \gamma_3 \exp(-\gamma_4 n) \\
&= \zeta_n(\epsilon).
\end{aligned}$$

□

Proof of Corollary 4.1. Given the explicit expression of ζ^t_n in Proposition A.1, Corollary 4.1 is the direct consequence of Theorem 4.1. □

A.4. Proof of Corollaries 4.2 and 4.3

To prove Corollaries 4.2 and 4.3, we need Propositions A.6 and A.7 respectively.

Proposition A.6. *Assume that Condition (T4) holds.*

(i) For the logistic regression model, there exists some positive constant α such that if $n^{\alpha/(\alpha+2)}\epsilon^2 \to \infty$, we have that $\text{pr}\left(\max_{\mathcal{J}} |\phi^{GLM}_{n\mathcal{J}} - \phi^{GLM}_{\mathcal{J}}| > \epsilon\right) \le (\prod_{r=1}^R p_r)\zeta^{logistic}_n(\epsilon)$, *where*

$$\zeta^{logistic}_n(\epsilon) = \gamma_1 \exp(-\gamma_2 n^{\alpha/(\alpha+2)}\epsilon^2) + \gamma_3 n \exp(-\gamma_4 n^{\alpha/(\alpha+2)}),$$

for some positive constants γ_1, γ_2, γ_3 and γ_4.

(ii) For linear models, there exists some positive constant α such that if $n^{\alpha/A}\epsilon^2 \to \infty$, we have that $\text{pr}\left(\max_{\mathcal{J}} |\phi^{GLM}_{n\mathcal{J}} - \phi^{GLM}_{\mathcal{J}}| > \epsilon\right) \le (\prod_{r=1}^R p_r)\zeta^{linear}_n(\epsilon)$, *where*

$$\zeta^{linear}_n(\epsilon) = \gamma_1 \exp\left(-\gamma_2 n^{\alpha/A}\epsilon^2\right),$$

for some positive constants γ_1, γ_2 and $A = \max(\alpha + 4, 3\alpha + 2)$.

Proposition A.7. *Under Condition (T4), for any $0 < v < 1/2$, we have that* $\text{pr}\left(\max_{\mathcal{J}} |\phi^{DC}_{n\mathcal{J}} - \phi^{DC}_{\mathcal{J}}| > \epsilon\right) \le (\prod_{r=1}^R p_r)\zeta^{DC}_n(\epsilon)$, *where*

$$\zeta^{DC}_n(\epsilon) = \gamma_1 \exp\left(-\gamma_2 n^{1-2v}\epsilon^2\right) + \gamma_3 n \exp(-\gamma_4 n^v),$$

for some positive constants γ_1, γ_2, γ_3 and γ_4.

Proposition A.6 is a straightforward extension of the main result in Theorem 4 in [13] to tensor case. Proposition A.7 is a straightforward extension of Theorem 1 in [25] to tensor case. The original theorems were developed for vectors. We simply rewrite them in tensor notations. Moreover, we replace $n^{-\kappa}$ with ϵ in the probability bounds. By Propositions A.6 and A.7, Condition (T4) implies Condition (T2).

Proof of Corollaries 4.2 and 4.3. Given the explicit expression of $\zeta_n^{logistic}$, ζ_n^{linear}, ζ_n^{DC} in Propositions A.6 and A.7, Corollaries 4.2 and 4.3 are the direct consequences of Theorem 4.1. □

A.5. Proof of Lemmas 4.2 and 4.3

We first present the following proposition and its proof. It is used to prove Lemma 4.2.

Proposition A.8. *Let $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ be a R-way random tensor and $\mathbf{X}$ follows a tensor normal distribution such that $\mathbf{X} \sim TN(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R)$. Let $\mathbf{M} \in \mathbb{R}^{d_1 \times \cdots \times d_R}$ be a constant tensor and $\mathbf{U}_r \in \mathbb{R}^{d_r \times p_r}, r = 1, \ldots, R$ be matrices with full row rank. If*

$$\mathbf{Y} = \mathbf{M} + \mathbf{X} \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R,$$

then $\mathbf{Y}$ follows a tensor normal distribution with parameters $(\mathbf{M}', \boldsymbol{\Sigma}_1', \ldots, \boldsymbol{\Sigma}_R')$ where $\mathbf{M}' = \mathbf{M} + \boldsymbol{\mu} \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R$ and $\boldsymbol{\Sigma}_r' = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{U}_r^T$ for $r = 1, \ldots, R$.

Proof of Proposition A.8. By definition, there exists a random tensor $\mathbf{Z}$ such that $\mathbf{Z} \sim TN(\mathbf{0}, \mathbf{I}_1, \ldots, \mathbf{I}_R)$ and

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{Z} \times_1 \boldsymbol{\Sigma}_1^{1/2} \cdots \times_R \boldsymbol{\Sigma}_R^{1/2}.$$

Therefore, we have

$$\begin{aligned}
\mathbf{Y} &= \mathbf{M} + \mathbf{X} \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R \\
&= \mathbf{M} + (\boldsymbol{\mu} + \mathbf{Z} \times_1 \boldsymbol{\Sigma}_1^{1/2} \cdots \times_R \boldsymbol{\Sigma}_R^{1/2}) \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R \\
&= \mathbf{M} + \boldsymbol{\mu} \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R + \mathbf{Z} \times_1 \boldsymbol{\Sigma}_1^{1/2} \cdots \times_R \boldsymbol{\Sigma}_R^{1/2} \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R \\
&= \mathbf{M} + \boldsymbol{\mu} \times_1 \mathbf{U}_1 \cdots \times_R \mathbf{U}_R + \mathbf{Z} \times_1 (\mathbf{U}_1 \boldsymbol{\Sigma}_1^{1/2}) \cdots \times_R (\mathbf{U}_R \boldsymbol{\Sigma}_R^{1/2})
\end{aligned}$$

where we use the property that $\mathbf{X} \times_r \mathbf{A} \times_r \mathbf{B} = \mathbf{X} \times_r (\mathbf{BA})$ from [21]. Since $\mathbf{U}_r \boldsymbol{\Sigma}_r^{1/2} (\mathbf{U}_r \boldsymbol{\Sigma}_r^{1/2})^T = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{U}_r^T = \boldsymbol{\Sigma}_r'$, we finish the proof. □

Proof of Lemma 4.2. In TDA model, we have

$$\mathbf{X} \mid (Y = k) \sim TN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_R).$$

By construction, $\mathbf{X}_{\mathcal{V}}$ is the sub-tensor of $\mathbf{X}$. Therefore, there exist some matrices $\mathbf{G}_1, \ldots, \mathbf{G}_R$ such that

$$\mathbf{X}_{\mathcal{V}} = \mathbf{X} \times_1 \mathbf{G}_1 \cdots \times_R \mathbf{G}_R$$

where $\mathbf{G}_r \in \mathbb{R}^{l_r \times p_r}$ and $l_r = |\mathcal{V}_r|$. Specifically, the ith row of $\mathbf{G}_r$ is the $\mathcal{V}_{r,i}$th standard basis $e_{\mathcal{V}_{r,i}}$ for the space $\mathbb{R}^{p_r}$ where $\mathcal{V}_{r,i}$ is the ith element in set $\mathcal{V}_r$. Therefore, $\mathbf{G}_r$ has full row rank. By Proposition A.8, we have

$$\mathbf{X}_{\mathcal{V}} \mid (Y = k) \sim TN\left(\boldsymbol{\mu}_k \times_1 \mathbf{G}_1 \cdots \times_R \mathbf{G}_R, \mathbf{G}_1\boldsymbol{\Sigma}_1\mathbf{G}_1^T, \ldots, \mathbf{G}_R\boldsymbol{\Sigma}_R\mathbf{G}_R^T\right).$$

Similar to the construction of $\mathbf{X}_{\mathcal{V}}$, we have $\boldsymbol{\mu}_k \times_1 \mathbf{G}_1 \cdots \times_R \mathbf{G}_R = \boldsymbol{\mu}_{k,\mathcal{V}}$. Moreover, the (i,j)th element in matrix $\mathbf{G}_r\boldsymbol{\Sigma}_r\mathbf{G}_r^T$ can be written as $e_{\mathcal{V}_{r,i}}^T\boldsymbol{\Sigma}_r e_{\mathcal{V}_{r,j}}$, which is exactly the (i,j)th element of matrix $\boldsymbol{\Sigma}_{r,\mathcal{V}}$. Therefore, we show that $\mathbf{G}_r\boldsymbol{\Sigma}_r\mathbf{G}_r^T = \boldsymbol{\Sigma}_{r,\mathcal{V}}$ and we complete the proof of Lemma 4.2. □

Proof of Lemma 4.3. In the rank-H generalized linear tensor regression model [59], we have

$$g(\mu) = \beta_0 + \langle \mathbf{B}, \mathbf{X} \rangle \tag{A.14}$$

where $g(\cdot)$ is the link function, $\mu = \mathrm{E}(Y \mid \mathbf{X})$ and β_0 is the intercept. In addition, the coefficient tensor $\mathbf{B} \in \mathbb{R}^{p_1 \times \ldots \times p_R}$ can be decomposed as

$$\mathbf{B} = \sum_{h=1}^{H} \mathbf{a}_h^1 \circ \cdots \circ \mathbf{a}_h^R$$

for $\mathbf{a}_h^r \in \mathbb{R}^{p_r}, h = 1, \ldots, H$. Elementwise,

$$B_{i_1,\ldots,i_R} = \sum_{h=1}^{H} a_{i_1 h}^1 \cdots a_{i_R h}^R \text{ for } i_r = 1, \ldots, p_r \text{ and } r = 1, \ldots, R,$$

where $a_{i_k h}^r$ is the i_kth element of a_h^r. Since $\mathcal{V} = \mathcal{V}_1 \times \cdots \times \mathcal{V}_R$, we let $[\mathbf{a}_h^r]_{\mathcal{V}_r}$ denote the sub-vector of $\mathbf{a}_h^r$ restricted on the index set $\mathcal{V}_r$, i.e. $[\mathbf{a}_h^r]_{\mathcal{V}_r} = (a_{ih}^r \mathbf{1}_{i \in \mathcal{V}_r})$. Similar to the construction of $\mathbf{X}_{\mathcal{V}}$, we define $\mathbf{B}_{\mathcal{V}}$. It is easy to see that

$$\mathbf{B}_{\mathcal{V}} = \sum_{h=1}^{H} [\mathbf{a}_h^1]_{\mathcal{V}_1} \circ \cdots \circ [\mathbf{a}_h^R]_{\mathcal{V}_R}.$$

Therefore, $\mathbf{B}_{\mathcal{V}}$ is also a tensor of rank-H. When $\mathcal{D} \subseteq \mathcal{V}$, $B_{\mathcal{J}} = 0$ for $\mathcal{J} \notin \mathcal{V}$. The tensor regression model (A.14) can be rewritten as

$$g(\mu) = \beta_0 + \langle \mathbf{B}_{\mathcal{V}}, \mathbf{X}_{\mathcal{V}} \rangle. \tag{A.15}$$

Thus, we can fit a rank-H generalized linear tensor regression model on $\mathbf{X}_{\mathcal{V}}$ and Y. □

Appendix B: Additional simulation results

We show additional simulation results for Models 1–3 and 8–9 using t-statistic and distance correlation in the screening procedure. Tables 7 and 8 report the classification errors on the original data and on the screened data with different screening weights. The results at $c = \omega/2$ and ω are generally similar, and they are better than the results at $c = 0$. This indicates that STS outperforms the marginal screening methods and is not overly sensitive to the choice of c.

TABLE 7

Classification errors for Models 1-3 and 8-9. STS-t is used for screening. Standard errors are reported in parentheses. For each model, the column $\mathbf{X}$ *corresponds to the results on the original dataset; the other three columns are results on data with different screening methods: marginal screening, and STS with two choices of c.*

	X	0	$\omega/2$	ω	X	0	$\omega/2$	ω	X	0	$\omega/2$	ω
	Model 1				Model 2				Model 3			
CATCH	20.04	19.41	15.78	14.87	18.11	17.60	13.91	13.67	13.17	11.58	9.57	9.59
	(0.27)	(0.20)	(0.17)	(0.15)	(0.18)	(0.19)	(0.14)	(0.12)	(0.15)	(0.16)	(0.10)	(0.09)
CP-GLM	21.87	21.50	21.36	21.38	20.95	21.58	19.40	18.93	19.81	14.03	12.93	12.66
	(0.70)	(0.29)	(0.29)	(0.33)	(0.38)	(0.32)	(0.24)	(0.25)	(1.26)	(0.40)	(0.27)	(0.23)
ℓ_1-GLM	23.10	22.70	17.38	16.33	20.15	19.42	15.70	15.51	15.69	14.34	12.52	12.80
	(0.32)	(0.33)	(0.19)	(0.16)	(0.19)	(0.23)	(0.14)	(0.14)	(0.19)	(0.18)	(0.14)	(0.14)
ℓ_1-FDA	31.95	20.89	14.88	13.73	18.77	15.99	12.31	12.28	16.08	11.86	10.16	10.64
	(0.36)	(0.26)	(0.11)	(0.09)	(0.24)	(0.17)	(0.09)	(0.07)	(0.30)	(0.16)	(0.08)	(0.09)
	Model 8				Model 9							
CATCH	3.30	3.26	4.23	14.18	18.98	18.88	16.95	16.74				
	(0.06)	(0.05)	(0.10)	(0.41)	(0.20)	(0.19)	(0.17)	(0.17)				
CP-GLM	26.17	20.82	20.85	26.62	23.53	22.35	19.40	19.11				
	(0.56)	(0.37)	(0.42)	(0.47)	(0.59)	(0.43)	(0.33)	(0.35)				
ℓ_1-GLM	4.82	4.61	5.35	22.37	20.97	20.66	17.99	18.46				
	(0.10)	(0.09)	(0.11)	(0.46)	(0.25)	(0.22)	(0.17)	(0.20)				
ℓ_1-FDA	3.18	3.30	4.21	22.45	26.89	19.94	16.62	17.03				
	(0.04)	(0.05)	(0.09)	(0.46)	(0.47)	(0.22)	(0.14)	(0.15)				

Appendix C: Effect of the neighborhood size

Define neighborhood size to be the number of elements in a neighborhood along each mode. The default STS proposed in Section 3.1 has a neighborhood size of 3. To study the effect of the neighborhood size in STS, we let the neighborhood size vary from 3 to 9. For all the models, we select the first $d_n = \lceil n/\log n \rceil$ predictors and compare the true positive rates (TPR) of the screened data.

We conducted our simulations with all methods on all models considered in Section 5.1. The pattern is similar across all models. For the sake of space, we only present the results for STS-GLM of two models, the smooth Model 1 and the non-smooth Model 8. The TPRs are plotted in Figure 5. Results show that, in the smooth Model 1, sizes of 5 and 7 give the best results, but the size of 3 is only slightly worse. If we take a large neighborhood size of 9, there is a notable decrease in TPR. On the other hand, in the non-smooth Model 8, the size of 3 with $c = \omega/2$ gives a reasonable TPR that is close to 80%, but the larger neighborhoods have drastically worse performance. Hence, unless there is a very strong belief in smoothness, a neighborhood size of 3 should be preferred.

Appendix D: Comparison with Gaussian filter

In this section, we compare STS with Gaussian filter, which is an image processing technique that blurs an image with the Gaussian function [16]. Gaussian

TABLE 8

Classification errors for Models 1-3 and 8-9. STS-DC is used for screening. Standard errors are reported in parentheses. For each model, the column $\mathbf{X}$ *corresponds to the results on the original dataset; the other three columns are results on data with different screening methods: marginal screening, and STS with two choices of* c*.*

	X	0	$\omega/2$	ω	X	0	$\omega/2$	ω	X	0	$\omega/2$	ω
	Model 1				Model 2				Model 3			
CATCH	20.04	19.33	15.41	14.25	18.11	17.46	13.79	13.62	13.17	11.60	9.43	9.50
	(0.27)	(0.26)	(0.20)	(0.12)	(0.18)	(0.18)	(0.13)	(0.12)	(0.15)	(0.14)	(0.08)	(0.08)
CP-GLM	21.87	21.34	21.68	20.62	20.95	22.05	19.43	18.66	19.81	14.03	12.22	12.84
	(0.70)	(0.35)	(0.32)	(0.31)	(0.38)	(0.40)	(0.28)	(0.22)	(1.26)	(0.30)	(0.22)	(0.31)
ℓ_1-GLM	23.10	22.69	17.31	15.96	23.31	19.46	15.70	15.55	21.62	14.48	12.14	12.39
	(0.32)	(0.33)	(0.20)	(0.16)	(0.31)	(0.20)	(0.14)	(0.14)	(0.31)	(0.19)	(0.13)	(0.12)
ℓ_1-FDA	31.95	20.79	14.65	13.56	26.59	16.15	12.29	12.33	32.37	12.16	9.39	9.81
	(0.36)	(0.25)	(0.12)	(0.09)	(0.42)	(0.17)	(0.08)	(0.07)	(0.70)	(0.16)	(0.08)	(0.08)
	Model 8				Model 9							
CATCH	3.30	3.27	4.03	15.95	18.98	18.87	16.50	16.08				
	(0.06)	(0.06)	(0.10)	(0.34)	(0.20)	(0.19)	(0.18)	(0.15)				
CP-GLM	26.17	20.69	19.75	26.89	23.53	22.71	18.85	18.64				
	(0.56)	(0.37)	(0.32)	(0.43)	(0.59)	(0.47)	(0.41)	(0.32)				
ℓ_1-GLM	4.82	4.62	5.22	20.98	20.97	20.65	17.61	17.61				
	(0.10)	(0.10)	(0.11)	(0.47)	(0.25)	(0.22)	(0.19)	(0.17)				
ℓ_1-FDA	3.18	3.31	3.84	20.76	26.89	19.79	16.11	16.05				
	(0.04)	(0.05)	(0.08)	(0.48)	(0.47)	(0.22)	(0.16)	(0.11)				

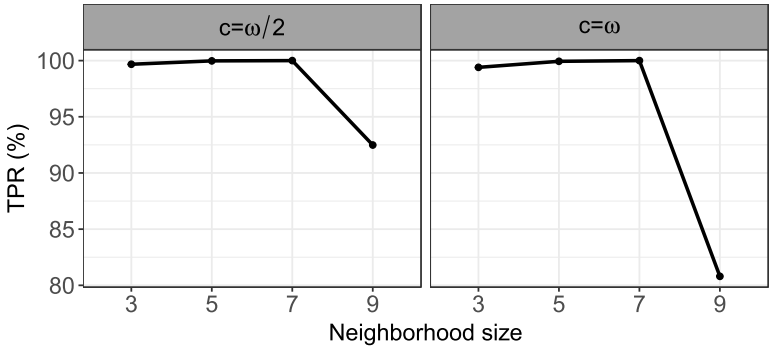
filter is usually used in a somewhat different context from ours. It is often applied on two-way or three-way tensors, and it smoothes images instead of screening utilities. However, it is straightforward to generalize it to our problem of interest. For easy presentation, we directly present a definition for R-way tensor screening. The Gaussian filter function for the R-dimensional space is defined as follows:

$$H(u_1, \ldots, u_R) = \exp\{-(u_1^2 + \cdots + u_R^2)/(2\eta^2)\}.$$

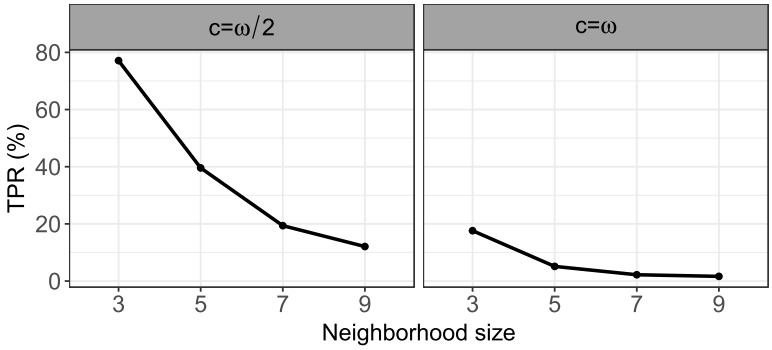
where u_r is the distance from the origin in the r-th mode and η is the standard deviation. Let d_{nb} be an odd number that denotes the size of the filter. Then u_r can take discrete values from the set $\mathcal{D}_{\text{nb}} = \{-(d_{\text{nb}}-1)/2, \ldots, -1, 0, 1, \ldots, (d_{\text{nb}}-1)/2\}$. Within the neighborhood, $H(u_1, \ldots, u_R)$ is normalized such that

$$H(u_1, \ldots, u_R) = H(u_1, \ldots, u_R) \Big/ \sum_{u_1, \ldots, u_R \in \mathcal{D}_{\text{nb}}} H(u_1, \ldots, u_R).$$

Then we have a R-way filter tensor which is also referred to as Gaussian kernel. After obtaining the kernel, we do a convolution between the kernel and the screening utility tensor. Let $\Phi \in \mathbb{R}^{p_1 \times \cdots \times p_R}$ denote the screening utility tensor where $\Phi_{j_1, \ldots, j_R}$ is the screening statistic corresponding to the predictor $X_{j_1, \ldots, j_R}$.



(a) Model 1



(b) Model 8

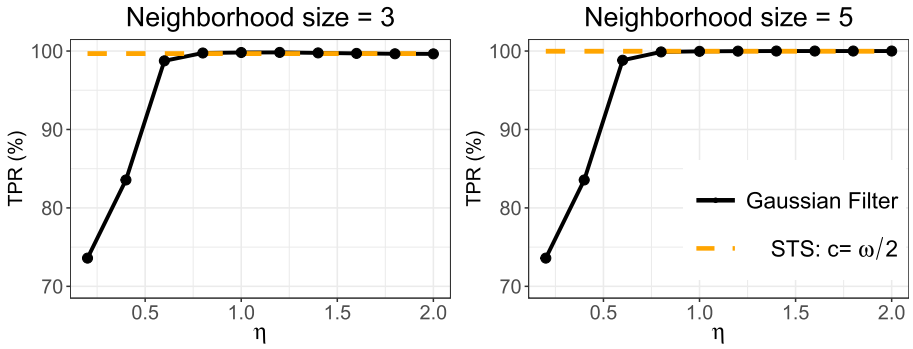
FIG 5. *TPR of STS-GLM for Models 1 & 8 with different neighborhood sizes.*

The filtered screening utility matrix Φ^{GF} is obtained by

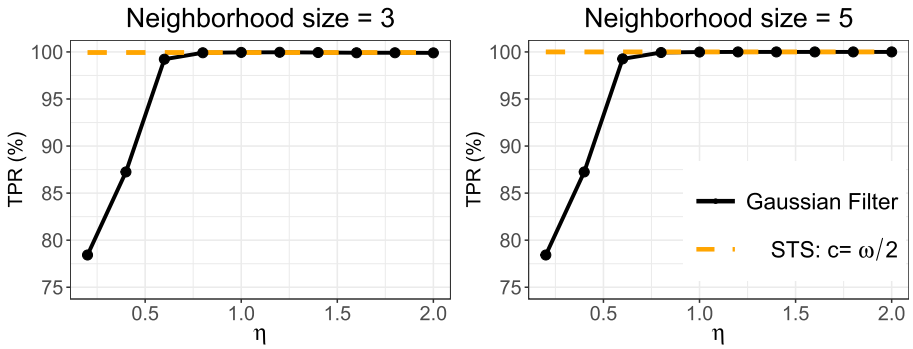
$$\Phi^{\mathrm{GF}}_{j_1,\ldots,j_R} = \sum_{u_1,\ldots,u_R \in \mathcal{D}_{\mathrm{nb}}} H(u_1,\ldots,u_R)\, \Phi_{j_1+u_1,\ldots,j_R+u_R}. \tag{D.1}$$

We will pad the tensor on the boundary if $j_r + u_r$ exceeds the index range. Then we will use $\Phi^{\mathrm{GF}}_{j_1,\ldots,j_R}$ to do screening.

We performed screening with Gaussian filter combined with the three screening utilities on all the models in Section 5.1. For all models, we select the first $d_n = \lceil n/\log n \rceil$ predictors and calculate the true positive rate (TPR). We consider two neighborhood sizes of $3, 5$ and a range of $\eta \in [0.2, 2]$. For the sake of space, only results for STS-GLM in the classification Model 1 and the regression Model 4 are shown in Figure 6. Other models exhibit the same pattern. Figure 6 shows that when the parameters in the Gaussian filter are chosen appropriately, it performs similarly to STS with $c = \omega/2$. It is particularly important to choose η in a reasonable range, but the screening results are not sensitive within this range. Note that the larger η is, the flatter is the Gaussian filter, and thus the



(a) Model 1



(b) Model 4

FIG 6. *Plots of TPR in screening for Models 1 & 4. Solid line and dashed line show the results of Gaussian filter and STS-GLM with* $c = \omega/2$*, respectively.*

smoother is the screening result. Figure 6 confirms that η should be somewhat large so that smoothness is encouraged to achieve better screening results.

Given the similar performance, we recommend using STS in practice unless there is strong prior knowledge that the Gaussian filter should be preferred. In general, the weighted average in STS is easier to interpret, and the role of c in STS can be more easily understood than the parameter η for researchers with relatively less background in statistics.

Acknowledgments

The authors thank the associate editor and referees, whose comments led to significant improvements of this paper. The authors are also grateful to Dr. Wen Li (Psychology, Florida State University) for helpful discussion.

References

[1] Brandi, G. and Di Matteo, T. (2021). Predicting multidimensional data via tensor learning. *Journal of Computational Science* **53** 101372. MR4249113

[2] Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106** 1566–1577. MR2896857

[3] Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics* **41** 2123–2148. MR3127860

[4] Chang, J., Tang, C. Y. and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics* **44** 515–539. MR3476608

[5] Chi, E. C. and Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications* **33** 1272–1299. MR3023474

[6] Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics* **53** 406–413. MR2850472

[7] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics* **36** 2605–2637. MR2485009

[8] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106** 544–557. MR2847969

[9] Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 745–771. MR2965958

[10] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911. MR2530322

[11] Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association* **109** 1270–1284. MR3265696

[12] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research* **10** 2013–2038. MR2550099

[13] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38** 3567–3604. MR2766861

[14] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.

[15] Gahrooei, M. R., Yan, H., Paynabar, K. and Shi, J. (2021). Multiple tensor-on-tensor regression: An approach for modeling processes with

heterogeneous sources of data. *Technometrics* **63** 147–159. MR4251490

[16] Gonzalez, R. C. and Woods, R. E. (2007). *Digital image processing (3rd Edition)*. Pearson.

[17] Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*. Chapman and Hall/CRC. MR1738933

[18] Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics* **18** 533–550. MR2751640

[19] He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41** 342–369. MR3059421

[20] Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics* **9** 1169–1193. MR3418719

[21] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review* **51** 455–500. MR2535056

[22] Kong, D., An, B., Zhang, J. and Zhu, H. (2020). L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association* **115** 403–424. MR4078472

[23] Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics* **38** 1094–1121. MR2604706

[24] Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112** 1131–1146. MR3735365

[25] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107** 1129–1139. MR3010900

[26] Liu, J., Zhu, C. and Liu, Y. (2020). Smooth compact tensor ring regression. *IEEE Transactions on Knowledge and Data Engineering* 1–1.

[27] Liu, J., Zhu, C., Long, Z., Huang, H. and Liu, Y. (2021). Low-rank tensor ring learning for multi-linear regression. *Pattern Recognition* **113** 107753.

[28] Liu, T., Yuan, M. and Zhao, H. (2017). Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition. *arXiv preprint arXiv:1702.07449*.

[29] Liu, Y., Liu, J. and Zhu, C. (2020). Low-rank tensor train coefficient array estimation for tensor-on-tensor regression. *IEEE Transactions on Neural Networks and Learning Systems* **31** 5402–5411. MR4189257

[30] Llosa-Vite, C. and Maitra, R. (2020). Reduced-rank tensor-on-tensor regression and tensor-variate analysis of variance. *arXiv preprint arXiv:2012.10249*.

[31] Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics* **27** 638–647. MR3863764

[32] Lyu, T., Lock, E. F. and Eberly, L. E. (2017). Discriminating sample groups with multi-way data. *Biostatistics* **18** 434–450. MR3824759

[33] Lyu, X., Sun, W. W., Wang, Z., Liu, H., Yang, J. and Cheng, G.

(2020). Tensor graphical model: Non-convex optimization and statistical inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42** 2024–2037.

[34] Mai, Q., Yang, Y. and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica* **29** 97–111. MR3889359

[35] Mai, Q. and Zou, H. (2012). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100** 229–234. MR3034336

[36] Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics* **43** 1471–1497. MR3357868

[37] Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661

[38] Michel, V., Eger, E., Keribin, C., Poline, J.-B. and Thirion, B. (2010). A supervised clustering approach for extracting predictive information from brain activation images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* 7–14. IEEE.

[39] Michel, V., Gramfort, A., Varoquaux, G., Eger, E. and Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. *IEEE Transactions on Medical Imaging* **30** 1328–1340.

[40] Min, K., Mai, Q. and Zhang, X. (2021). Fast and separable estimation in high-dimensional tensor Gaussian graphical models. *Journal of Computational and Graphical Statistics* **0** 1–7.

[41] Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2007). Scan clustering: A false discovery approach. *Journal of Multivariate Analysis* **98** 1441–1469. MR2364129

[42] Pan, Y., Mai, Q. and Zhang, X. (2019). Covariate-adjusted tensor classification in high dimensions. *Journal of the American statistical association* **114** 1305–1319. MR4011781

[43] Raskutti, G. and Yuan, M. (2015). Convex regularization for high-dimensional tensor regression. *arXiv preprint arXiv:1512.01215*.

[44] Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics* **37** 2922–2952. MR2541451

[45] Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics* **39** 1241–1265. MR2816353

[46] Sharbrough, F., Chatrian, G. E., Lesser, R., Luders, H., Nuwer, M. and Picton, T. (1991). American electroencephalographic society guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology* **8** 200–202.

[47] Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* **6** 174–215.

[48] Sun, W. W. and Li, L. (2017). STORE: Sparse tensor response regression

and neuroimaging analysis. *The Journal of Machine Learning Research* **18** 4908–4944. MR3763769

[49] Sun, W. W. and Li, L. (2019). Dynamic tensor clustering. *Journal of the American Statistical Association* **114** 1894–1907. MR4047308

[50] Sun, W. W., Lu, J., Liu, H. and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 899–916. MR3641413

[51] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794. MR2382665

[52] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. MR3967104

[53] Wang, X., Zhu, H. and Initiative, A. D. N. (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association* **112** 1156–1168. MR3735367

[54] Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 753–772. MR2867457

[55] Xu, P., Zhu, J., Zhu, L. and Li, Y. (2015). Covariance-enhanced discriminant analysis. *Biometrica* **102** 33–45. MR3335094

[56] 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. MR4047294

[57] Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics* **59** 426–436. MR3740960

[58] Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin* **38** 531–538.

[59] Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108** 540–552. MR3174640

[60] Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106** 1464–1475. MR2896849