Statistica Sinica Preprint No: SS-2024-0281	
Title	Generalized Tensor Regression with Internal Variation Regularization
Manuscript ID	SS-2024-0281
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0281
Complete List of Authors	Yang Bai, Ting Li and Yang Sui
Corresponding Authors	Yang Bai
E-mails	statbyang@mail.shufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

Generalized Tensor Regression with Internal Variation Regularization

Yang Bai, Ting Li, and Yang Sui

Shanghai University of Finance and Economics

Abstract: Analyzing clinical diagnosis along with high-dimensional imaging data, while accounting for the piecewise constant nature of the imaging, presents challenges to existing statistical approaches. In this paper, we propose a generalized tensor regression framework with Internal Variation (IV) regularization to address these challenges. The inclusion of IV regularization allows for the explicit utilization of the rich spatial structure, particularly the piecewise constant nature of high-order imaging data, albeit with a more complex algorithm and demanding theoretical investigation. We develop an efficient IV regularized optimization procedure for estimating unknown scalar and tensor coefficients. We investigate the theoretical properties of scalar and tensor coefficient estimates, especially the error bounds of regularized tensor coefficient estimates. Extensive numerical studies assess the finite sample performance of our method, demonstrating its superiority over existing approaches. Finally, we apply the proposed method to a chronic sinusitis computed tomography (CT) imaging dataset and identify the most activated subregion across the maxillary sinus cavity associated with the diagnosis.

All authors contributed equally to this paper, and their names are listed in alphabetical order.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

Key words and phrases: CP decomposition, internal variation, piecewise constant, regularized GLM, tensor regression.

1. Introduction

Tensor data, represented as multi-dimensional arrays, naturally arise in various fields such as imaging, neuroscience, and spatiotemporal analysis, where the data exhibit inherent structural dependencies. Tensor regression extends traditional regression models by leveraging these multi-way structures to capture complex relationships between tensor predictors and scalar or tensor outcomes, with widespread applications across various fields (Bi et al., 2021; Liu et al., 2022). In image processing, tensor regression has been used for tasks such as denoising (Zhang et al., 2021), medical diagnosis (Zhou et al., 2013; Li and Zhang, 2021; Feng et al., 2021), and brain connectivity analysis (Spencer et al., 2019). In addition, it plays a crucial role in examining the relationship between manufacturing parameters and the geometry of manufactured parts (Yan et al., 2019). In multi-task learning, tensor regression uses shared information across tasks, enhancing model accuracy and outperforming independent task learning models (Yang and Hospedales, 2017). Furthermore, in spatio-temporal analysis, tensor regression has been extended to address both forecasting and cokriging tasks (Yu et al., 2018; Su et al., 2020).

In tensor regression, incorporating tensor predictors presents several challenges in the estimation and theoretical investigation. One key chal-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

lenge is handling the multi-dimensional tensor data that possess complex structures dependencies. Straightforward approaches, such as vectorizing tensors and treating the resulting vectors as covariates (Zhou et al., 2014), lead to ultrahigh-dimensional models that impose significant computational burdens and disregard the spatial organization of the data. To address the complexity of tensor structures, various low-dimensional structural assumptions for tensors have been proposed, including element-wise, fiber-wise, or slice-wise sparsity (Raskutti et al., 2019; Zhang et al., 2019) and low-rankness (Raskutti et al., 2019; Luo and Zhang, 2024). (Raskutti et al., 2019) systematically studied these sparsity and low-rank structures, establishing general risk bounds and specific upper bounds across different scenarios. Another line of research has focused on tensor regression models leveraging tensor decompositions to exploit the high-order structure of tensor data, such as Canonical Polyadic (CP) and Tucker decompositions (Kolda and Bader, 2009), with theoretical guarantees established in (Zhou et al., 2013; Li et al., 2018; Lu et al., 2020; Wu and Feng, 2023). While these methods effectively utilize the structural properties of the tensor data, they overlook the inherent structure of the corresponding tensor coefficients.

Recovering true tensor coefficients while preserving structural information is crucial in practical applications to enhance interpretability. This challenge is particularly significant in high-dimensional imaging studies, where tensor coefficients are expected to exhibit piecewise constant patterns

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

(Feng et al., 2021; Li and Zhang, 2021), and neighboring voxels tend to have similar coefficient values (Michel et al., 2011; Wang et al., 2017). Effective regularization methods are required to exploit these spatial dependencies while maintaining computational efficiency. Wang et al. (2017) proposed a two-dimensional total variation (TV) method to enforce piecewise constant structures in matrix coefficients. However, due to its computational complexity, this approach does not scale well to third-order or higher-order tensor images. To address this limitation, Feng et al. (2021) introduced Internal Variation (IV) regularization, an extension of anisotropic TV to higher-order tensors for linear tensor regression. While IV regularization provides a more scalable solution, it is restricted to continuous response variables and is not applicable to discrete outcomes. Moreover, its theoretical properties remain unexplored.

To the best of our knowledge, inference for the piecewise constant nature of imaging coefficients in generalized tensor regression models has not been studied in the literature. We aim to develop a generalized tensor regression model tailored for high-dimensional imaging data, while providing theoretical guarantees.

In this paper, we develop a generalized tensor regression model including scalar and tensor predictors to cover different distributions of responses, and apply IV regularization to address the piecewise constant nature of tensor coefficients. Our approach makes several key contributions. First,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

we impose the CP decomposition to reduce the dimension of tensor coefficients while preserving the spatial structure of the imaging data. We then construct IV regularization based on the CP decomposition, effectively capturing the piecewise constant characteristics of imaging coefficients. Unlike conventional fusion penalties (Li and Zhang, 2021), which apply a uniform regularization strength across all factor vectors and overlook those with smaller total variations, IV regularization dynamically adjusts regularization strengths across different factor vectors. This adaptive approach allows for accurate estimation of both high- and low-variation components, improving both interpretability and estimation precision. Second, we extend the Alternating Direction Method of Multipliers (ADMM) to iteratively update CP decomposition coefficients under IV regularization. Existing ADMM-based tensor regression methods (Lu et al., 2020; Li and Zhang, 2021) cannot be directly applied due to the additional complexity introduced by the IV penalty and the nontrivial likelihood function. Our modified ADMM framework overcomes these computational challenges, enabling efficient and scalable estimation. Third, we introduce a novel bootstrap procedure specifically designed for tensor imaging data to assess the significance of both scalar and tensor coefficients, with a particular focus on specific subregions of the tensor coefficient. This addresses a critical gap in the literature, as no existing inference methods have been developed for tensor regression models.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

In theory, we investigate the theoretical properties of both scalar and tensor coefficient estimates in non-regularized and regularized tensor regression. Specifically, we first establish the asymptotic normality of these estimates in the non-regularized setting. We then derive an error bound for the coefficient estimates under general regularization and show that with appropriately diminishing regularization, the estimated coefficients converge to the true values as the sample size approaches infinity. Furthermore, we establish an error bound for the tensor coefficient estimate under IV regularization. Notably, we show that this error bound is related to the sparsity in the internal variation of the factor vectors in CP decompositions, highlighting the impact of the piecewise constant nature of high-order imaging data on estimation accuracy. Due to the complexity of IV regularization, deriving theoretical guarantees for this method is challenging. Our proof involves technical considerations linking the internal variation of the decomposed factor vectors with the boundedness of the tensor coefficients beyond those typically required for generalized tensor regression with simpler penalties like the fusion penalty. To the best of our knowledge, this is the first paper to theoretically explore IV regularization for the piecewise constant nature of high-order imaging data.

Finally, we apply the proposed method to a chronic sinusitis (CRS) CT imaging dataset, identifying subregions of the maxillary sinus cavity associated with pathogenesis, which have significant biological implications.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

Although there has been extensive work linking brain imaging data to brain disorders such as Alzheimer's disease (Jack et al., 2010) and ADHD (Hinshaw and Scheffler, 2014), research on imaging data related to other body parts is still very scarce.

The rest of the article is organized as follows. In Section 2, we introduce the generalized tensor regression model based on CP decomposition, and IV. Section 3 and Section 4 present the estimation and implementation details and the theoretical properties, respectively. In Section 5, we provide the simulation studies for 3D scenarios. Section 6 illustrates the CRS application. The last section provides concluding remarks and discussion.

2. Methodology

2.1 Notation and Operations

We start with a brief summary of tensor notation and some array operations. Extensive references can be found in the review article (Kolda and Bader, 2009). Throughout this paper, we denote tensors by boldface script capital letters such as $\mathcal{X}, \mathcal{Y}$, matrices by boldface capital letters $\boldsymbol{X}, \boldsymbol{Y}$, vectors by small boldface letters $\boldsymbol{x}, \boldsymbol{y}$, and scalars by small letters x, y. A Dth-way tensor refers to a D-dimensional array $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$, where the dimension D of a tensor is known as modes and p_d $(1 \leq d \leq D)$ is the marginal dimension of the dth mode. For a vector $\boldsymbol{x}$, let $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$ denote its ℓ_1, ℓ_2 and ℓ_∞ norms. For a matrix $\boldsymbol{X}$, its Frobenius norm and vectorization are denoted by $\|\boldsymbol{X}\|_F$ and $\text{vec}(\boldsymbol{X})$. The matricization of a

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

tensor links the concepts and properties of matrices to tensors. The mode-d martricization of $\mathcal{X}$, denoted as $\mathcal{X}_{(d)}$ is defined as a $p_d \times \prod_{d' \neq d} p_{d'}$ matrix such that the $(i_1, \ldots, i_D)$th element of the tensor $\mathcal{X}$ maps to the (i_d, j)th element of the matrix $\mathcal{X}_{(d)}$, where $j = 1 + \sum_{d' \neq d}(i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$ (Kolda and Bader, 2009). Moreover, let $\text{vec}(\mathcal{X})$ represent the vectorization of a tensor $\mathcal{X}$, which stacks its elements into a columns vector with length $\prod_d^D p_d$ with its the jth entry maps to the $(i_1, \ldots, i_D)$th element of $\mathcal{X}$, where $j = 1 + \sum_{d=1}^D (i_d - 1) \prod_{d'=1}^{d-1} p_{d'}$. Next, we introduce some useful operations between tensors and matrices. Given $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q] \in \mathbb{R}^{p \times q}$, the kronecker product is the $mp \times nq$ matrix $\boldsymbol{A} \otimes \boldsymbol{B} = [\boldsymbol{a}_1 \otimes \boldsymbol{B} \ldots \boldsymbol{a}_n \otimes \boldsymbol{B}]$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of columns $n = q$, then the Khatri-Rao product is defined as an $mp \times n$ matrix by $\boldsymbol{A} \odot \boldsymbol{B} = [\boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \ldots \boldsymbol{a}_n \otimes \boldsymbol{b}_n]$. Next, we define the inner product $\langle \cdot, \cdot \rangle$ of two tensors with the same dimensions as $\langle \mathcal{X}, \mathcal{Y} \rangle = \langle \text{vec}(\mathcal{X}) \text{vec}(\mathcal{Y}) \rangle$ and the Frobenius norm of a tensor is defined as $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$. An outer product $\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \ldots \circ \boldsymbol{b}_D$ of D vectors $\boldsymbol{b}_1 \in \mathbb{R}^{p_1}, \ldots, \boldsymbol{b}_D \in \mathbb{R}^{p_D}$ is an array of dimension $p_1 \times \ldots \times p_D$ with its $(i_1, \ldots, i_D)$th element equal to $\prod_{d=1}^D \boldsymbol{b}_{d, i_d}$.

2.2 Model framework

In this paper, we consider a broader family of exponential distributions for the scalar response, which covers more types of tensor regression models. We assume that the response y given covariate $\boldsymbol{z} \in \mathbb{R}^q$ and tensor predictor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$ belongs to an exponential family with probability mass

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

function or density

$$p(y|\theta,\phi) = \exp\left\{\frac{y\theta - \psi(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where θ and $\phi > 0$ are the natural and dispersion parameters. The density is related to $\boldsymbol{z}$ and $\mathcal{X}$ through the linear systematic part given by where θ and $\phi > 0$ are the natural and dispersion parameters. The density is related to $\boldsymbol{z}$ and $\mathcal{X}$ through the linear systematic part given by where θ and $\phi > 0$ are the natural and dispersion parameters. The density is related to $\boldsymbol{z}$ and $\mathcal{X}$ through the linear systematic part given by

$$g(\mu) = \boldsymbol{z}^T\boldsymbol{\kappa} + \langle \mathcal{X}, \mathcal{A} \rangle, \tag{2.1}$$

where $\mu = E(y|\boldsymbol{z}, \mathcal{X})$ and $g(\cdot)$ is an increasing link function. $\boldsymbol{\kappa} \in \mathbb{R}^q$ and $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$ are the covariate and tensor coefficients corresponding to $\boldsymbol{z}$ and $\mathcal{X}$, respectively. It is clear that the number of parameters to be estimated in (2.1) is $q + \prod_{d=1}^{D} p_d$. This number can be very large, even for small values of each p_d. For example, a conventional MRI image with a pixel size of $256 \times 256 \times 256$ requires $256^3 = 16,777,216$ parameters, which is ultrahigh dimensional and exceeds the usual sample size. Including tensor predictors introduces computational burdens and difficulties in estimation.

To deal with the high dimensionality resulting from $\mathcal{A}$, we employ tensor decomposition for dimension reduction. Common decomposition methods include CP decomposition and Tucker decomposition (Kolda and Bader,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

2009). We opt for CP decomposition because, in general, the number of parameters obtained by CP decomposition is less than that of Tucker decomposition, and Tucker decomposition requires identifying a specific rank or size along each dimension (Li et al., 2018). For CP decomposition, a rank-R approximation of $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$ is presented as

$$\mathcal{A} = \sum_{r=1}^{R} \boldsymbol{a}_1^r \circ \boldsymbol{a}_2^r \circ \cdots \circ \boldsymbol{a}_D^r, \tag{2.2}$$

where $\boldsymbol{a}_d^r \in \mathbb{R}^{p_d}$, a column vector, corresponds to the dth mode and rth component, and R is defined as the rank of $\mathcal{A}$. The CP decomposition can be expressed as $\mathcal{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_D]$, where $\boldsymbol{A}_d = [\boldsymbol{a}_d^1, \boldsymbol{a}_d^2, \ldots, \boldsymbol{a}_d^R] \in \mathbb{R}^{p_d \times R}$ is the factor matrix of the d-th mode. In (2.2), potential identifiability issues can be attributed to scaling, permutation, and non-uniqueness of the decomposition. Constrained parameterization methods are needed to deal with the identifiability issues. Following Zhou et al. (2013); Li et al. (2018), we scale $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{D-1}$ such that $a_{d,1}^r = 1$ and arrange the first row entries of $\boldsymbol{A}_D$ in descending order $a_{D,1}^1 > \cdots > a_{D,1}^R$ to deal with the complication. After eliminating the indeterminacies, model (2.1) reduces to

$$g(\mu) = \boldsymbol{z}^T \boldsymbol{\kappa} + \left\langle \mathcal{X}, \sum_{r=1}^{R} \boldsymbol{a}_1^r \circ \boldsymbol{a}_2^r \circ \cdots \circ \boldsymbol{a}_D^r \right\rangle. \tag{2.3}$$

In comparison to model (2.1), model (2.3) not only preserves the complex spatial structure of the tensor data but also significantly reduces the number of parameters from $\prod_d^D p_d$ to $R \sum_{d=1}^D p_d$. In addition to a massive reduc-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

tion in dimensionality, the CP decomposition also provides a reasonable approximation to many low-rank signals.

2.3 IV regularization

An important feature of imaging coefficients is their piecewise constant nature, such that voxels in close proximity are more likely to have similar coefficients. A common strategy to capture this property in 2D images involves controlling the Total Variation (TV) (Wang et al., 2017). However, when it comes to higher-dimensional imaging data, the direct application of TV faces substantial computational and analytical challenges due to the increased complexity. To navigate these challenges in higher-dimensional data, the notion of Internal Variation (IV) was introduced in linear tensor regression (Feng et al., 2021). This concept assumes that the tensor coefficients $\mathcal{A}$ can be approximated by a rank-R CP decomposition. Thus, the IV of $\mathcal{A}$ is defined as

$$\|\mathcal{A}\|_{\mathrm{IV}} = \sum_{r=1}^{R} \prod_{d=1}^{D} \|\boldsymbol{a}_d^r\|_{\mathrm{TV}}, \tag{2.4}$$

where $\|\cdot\|_{\mathrm{TV}}$ for any $\boldsymbol{a} \in \mathbb{R}^p$ represents the TV of the vector: $\|\boldsymbol{a}\|_{\mathrm{TV}} = \sum_{i=2}^{p} |a_i - a_{i-1}|$. In a similar spirit to Feng et al. (2021), we consider an IV regularized approach to capture the piecewise constant nature of high-order imaging coefficients in generalized linear models. To the best of our knowledge, this is the first time that piecewise constant tensor coefficients are considered in generalized tensor regression models.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

We discuss in detail the advantages of the construction of IV regularization. As illustrated in (2.4), the essence of IV regularization lies in imposing a constraint on the total variations of all vectors $\{\boldsymbol{a}_d^r\}_{r,d}$ and assigning different regularization strengths to all $\{\|\boldsymbol{a}_d^r\|_{\mathrm{TV}}\}_{r,d}$. For example, $\|\boldsymbol{a}_d^r\|_{\mathrm{TV}}$ is penalized by $\lambda_n \prod_{d' \neq d} \|\boldsymbol{a}_{d'}^r\|_{\mathrm{TV}}$. This differs from the common fusion penalty $\|\mathcal{A}\|_{\mathrm{fusion}} = \sum_{r=1}^{R} \sum_{d=1}^{D} \|\boldsymbol{R}_d \boldsymbol{a}_d^r\|_1$ with the 1st-order differencing matrix $\boldsymbol{R}_d$ (Li and Zhang, 2021), where all $\{\|\boldsymbol{a}_d^r\|_{\mathrm{TV}}\}_{r,d}$ are penalized with the same strength λ_n. This differential regularization strength is crucial in estimating high-dimensional image coefficients with complex spatial structures, as it prioritizes smaller total variations $\|\boldsymbol{a}_d^r\|_{\mathrm{TV}}$ while ensuring accurate estimation of $\boldsymbol{a}_d^r$ with larger total variations. In contrast, with the fusion penalty, the penalty for smaller total variations $\|\boldsymbol{a}_d^r\|_{\mathrm{TV}}$ may be overshadowed and hence neglected due to the dominance of larger total variations.

To illustrate this, we give a toy example. Let $R = 1$, $D = 3$, and $\mathcal{A}_0 = \boldsymbol{a}_1 \circ \boldsymbol{a}_2 \circ \boldsymbol{a}_3$, where $\boldsymbol{a}_1$ is a 30-dimensional vector whose 2nd and 3rd elements are equal to 2, $\boldsymbol{a}_2$ is a 30-dimensional vector whose 2nd and 3rd elements are equal to 1, and $\boldsymbol{a}_3$ is a 30-dimensional vector whose 2nd and 3rd elements are equal to 0.5. This also implies that $\|\boldsymbol{a}_1\|_{\mathrm{TV}} \geq \|\boldsymbol{a}_2\|_{\mathrm{TV}} \geq \|\boldsymbol{a}_3\|_{\mathrm{TV}}$. After generating the binary response y as in Section 5, we estimate $\widehat{\boldsymbol{a}}_1$, $\widehat{\boldsymbol{a}}_2$ and $\widehat{\boldsymbol{a}}_3$ using the IV and fusion regularizations with the same strength λ_n for both methods. To ensure identifiability, we scale the second elements of $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ to 2 and 1, respectively. Figure 1 shows the trends of each element of $\widehat{\boldsymbol{a}}_1$,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

$\widehat{\boldsymbol{a}}_2$, and $\widehat{\boldsymbol{a}}_3$ with the number of iterations. We observe that the IV method converges faster and exhibits smaller fluctuations. Notably, the 2nd and 3rd elements of $\widehat{\boldsymbol{a}}_3$ estimated using IV regularization are always very close, whereas there is a significant difference between the two elements estimated using the fusion penalty. Additionally, the estimates of $\widehat{\boldsymbol{a}}_1$ and $\widehat{\boldsymbol{a}}_2$ obtained using the IV method are more accurate. This is because in the fusion penalty, the regularization strength of $\|\boldsymbol{a}_1\|_{\mathrm{TV}}$, $\|\boldsymbol{a}_2\|_{\mathrm{TV}}$, and $\|\boldsymbol{a}_3\|_{\mathrm{TV}}$, are the same, and $\|\mathcal{A}\|_{\mathrm{fusion}}$ is mainly affected by $\|\boldsymbol{a}_1\|_{\mathrm{TV}}$ and $\|\boldsymbol{a}_2\|_{\mathrm{TV}}$, which results in $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, which have larger total variations, being accurately estimated, while $\boldsymbol{a}_3$, which has smaller total variation, being ignored. On the contrary, in $\|\mathcal{A}\|_{\mathrm{IV}}$, the regularization strength for $\boldsymbol{a}_3$ is $\lambda_n\|\boldsymbol{a}_1\|_{\mathrm{TV}}\|\boldsymbol{a}_2\|_{\mathrm{TV}}$, and such a regularization strategy not only ensures that $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are accurately estimated, but also adequately takes into account the smaller $\|\boldsymbol{a}_3\|_{\mathrm{TV}}$.

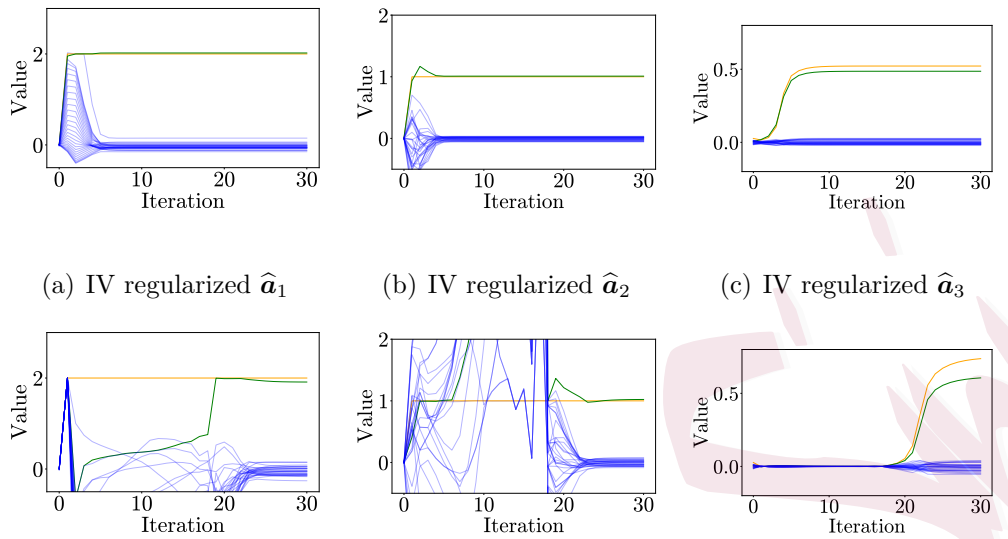
3. Estimation

3.1 Regularized GLM with IV penalty

Given the dataset $\{y_i, \boldsymbol{z}_i, \mathcal{X}_i\}_{i=1}^n$, the negative log-likelihood function is $L(\boldsymbol{\kappa}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_D) = \sum_{i=1}^n l\{(\boldsymbol{\kappa}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_D), y_i\} = -\sum_{i=1}^n [\{y_i\theta_i - \psi(\theta_i)\}/a(\phi) + c(y_i, \phi)]$, where θ_i is related to parameters $(\boldsymbol{\kappa}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_D)$ through (2.3) and $l(\cdot, y)$ denotes the negative log-likelihood. We consider the following IV regularized minimization problem

$$(\widehat{\boldsymbol{\kappa}}, \widehat{\boldsymbol{A}}_1, \ldots, \widehat{\boldsymbol{A}}_D) = \underset{\boldsymbol{\kappa}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_D}{\operatorname{argmin}} L(\boldsymbol{\kappa}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_D) + \lambda_n \|\mathcal{A}\|_{\mathrm{IV}}, \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281



(a) IV regularized $\widehat{\boldsymbol{a}}_1$ (b) IV regularized $\widehat{\boldsymbol{a}}_2$ (c) IV regularized $\widehat{\boldsymbol{a}}_3$

(d) Fusion regularized $\widehat{\boldsymbol{a}}_1$ (e) Fusion regularized $\widehat{\boldsymbol{a}}_2$ (f) Fusion regularized $\widehat{\boldsymbol{a}}_3$

Figure 1: Trends of each element in $\widehat{\boldsymbol{a}}_1$, $\widehat{\boldsymbol{a}}_2$ and $\widehat{\boldsymbol{a}}_3$ over iterations using the IV penalty and fusion penalty. The orange line represents the 2nd element and the blue line represents the 3rd element.

where λ_n is the regularization parameter. The estimation problem (3.1) can follow a block relaxation algorithm. However, deriving a piecewise constant estimate through the IV regularization comes at the cost of a more complex algorithm. Specifically, the form of the TV product, as in (2.4), across different modes poses significant challenges for estimation. We consider converting the IV regularization into a generalized lasso with the form $\|\boldsymbol{D}\boldsymbol{a}\|_1$ to remove the concatenated product form, where the generalized matrix $\boldsymbol{D}$ does not contain any parameters of $\boldsymbol{a}$. We next illustrate this idea in detail.

Let $\boldsymbol{a}_d = \text{vec}(\boldsymbol{A}_d) \in \mathbb{R}^{Rp_d}$ be the vectorization of $\boldsymbol{A}_d$. When updating

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

$\boldsymbol{a}_d$, the inner product in (2.3) can be rewritten as

$$\left\langle \mathcal{X}_i, \sum_{r=1}^{R} \boldsymbol{a}_1^r \circ \boldsymbol{a}_2^r \cdots \circ \boldsymbol{a}_D^r, \right\rangle = \sum_{r=1}^{R} \mathcal{X}_{i,(d)} \left(\boldsymbol{a}_1^r \otimes \cdots \otimes \boldsymbol{a}_{d-1}^r \otimes \boldsymbol{a}_{d+1}^r \otimes \cdots \otimes \boldsymbol{a}_D^r \right) \boldsymbol{a}_d^r. \tag{3.2}$$

Let $\boldsymbol{x}_{i,d}^r = \mathcal{X}_{i,(d)} \left(\boldsymbol{a}_1^r \otimes \ldots \otimes \boldsymbol{a}_{d-1}^r \otimes \boldsymbol{a}_{d+1}^r \otimes \ldots \otimes \boldsymbol{a}_D^r \right)$ and $\boldsymbol{v}_r$ be an R-dimensional vector with the rth element being 1 and the rest being 0. We have

$$\boldsymbol{X}_{i,d} = \left(\sum_{r=1}^{R} \boldsymbol{v}_r \otimes \boldsymbol{x}_{i,d}^r \right) \in \mathbb{R}^{Rp_d}, \quad \boldsymbol{X}_d = [\boldsymbol{X}_{1,d}, \ldots, \boldsymbol{X}_{n,d}] \in \mathbb{R}^{n \times Rp_d}. \tag{3.3}$$

As a result, the expression in (3.2) takes the form of $\boldsymbol{X}_d \boldsymbol{a}_d$ for each d. This implies that by fixing other $\boldsymbol{A}_{d' \neq d}$ matrices, we can update each $\boldsymbol{a}_d$ through traditional GLM regression with Rp_d parameters.

For the IV regularization, notice that $\|\boldsymbol{a}_d^r\|_{\mathrm{TV}} = \|\boldsymbol{G}_d \boldsymbol{a}_d^r\|_1$ where $\boldsymbol{G}_d \in \mathbb{R}^{(p_d-1) \times p_d}$ is the 1st-order differencing matrix. Let $\xi_d^r = \prod_{d' \neq d} \|\boldsymbol{a}_{d'}^r\|_{\mathrm{TV}}$ and $\boldsymbol{H}_r$ be an $R \times R$ identity matrix , then

$$\|\mathcal{A}\|_{\mathrm{IV}} = \sum_{r=1}^{R} \left(\prod_{d' \neq d} \|\boldsymbol{a}_{d'}^r\|_{\mathrm{TV}} \right) \|\boldsymbol{a}_d^r\|_{\mathrm{TV}} = \|\boldsymbol{D}_d \boldsymbol{a}_d\|_1, \tag{3.4}$$

where $\boldsymbol{D}_d = \sum_{r=1}^{R} \xi_d^r \boldsymbol{H}_r \otimes \boldsymbol{G}_d \in \mathbb{R}^{R(p_d-1) \times Rp_d}$. By combining (3.2), (3.3), and (3.4), we can estimate $\boldsymbol{a}_d$, given $\boldsymbol{\kappa}$ and $\boldsymbol{A}_{d' \neq d}$, through a generalized lasso regularized GLM:

$$\widehat{\boldsymbol{a}}_d = \underset{\boldsymbol{a}_d}{\operatorname{argmin}} \, L(\boldsymbol{X}_d \boldsymbol{a}_d) + \lambda_n \|\boldsymbol{D}_d \boldsymbol{a}_d\|_1. \tag{3.5}$$

For the generalized lasso optimization (3.5), note that the non-smoothness of the $\|\boldsymbol{D}_d \boldsymbol{a}_d\|_1$ makes traditional algorithms, such as gradient descent, suffer from slow convergence. We adopt the Alternating Directional Method

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

of Multipliers (ADMM) algorithm, which has been frequently used in distributed settings (Boyd et al., 2011). The detailed information about the ADMM algorithm is included in the Supplement Material S.1 to save space. The $\boldsymbol{a}_d$'s are recursively updated by fixing other estimates until convergence. The steps of the IV regularized generalized tensor regression are summarized in Algorithm 1 in the Supplement Material S.2. In practice, the initial values of $(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_D)$ are chosen from independent Normal distributions with small variance. We select the tuning parameters λ_n and R (if unknown) using the following Bayesian information criterion (BIC), as described in Burnham and Anderson (2004); Feng et al. (2021),

$$\mathrm{BIC}(\lambda_n, R) = 2L(\widehat{\boldsymbol{\kappa}}, \widehat{\mathcal{A}}) + \log n \cdot df(\lambda, R),$$

where $df(\lambda, R)$ represents the degrees of freedom. We use the degrees of freedom defined in Feng et al. (2021) and Li and Zhang (2021), which equals the sum of the nonzero elements across $\{\boldsymbol{D}_d \boldsymbol{a}_d\}_d$. We provide a more detailed discussion on the BIC in the Supplement Material S.10.

3.2 Hypothesis Testing

Although the estimation of parameters is interesting, it is equally important to test for the significance of $\boldsymbol{\kappa}$ and $\mathrm{vec}(\mathcal{A})$. This is particularly relevant in biomedical imaging analysis, where we aim to examine whether imaging data and other covariates have significant effects on diagnosis. Therefore,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

we propose to test two sets of null and alternative hypotheses as follows:

$$H_{0,\kappa} : \boldsymbol{\kappa} = \mathbf{0}_q, \quad vs. \quad H_{1,\kappa} : \boldsymbol{\kappa} \neq \mathbf{0}_q,$$

$$H_{0,\mathcal{A}} : \mathrm{vec}(\mathcal{A}) = \mathbf{0}_p, \quad vs. \quad H_{1,\mathcal{A}} : \mathrm{vec}(\mathcal{A}) \neq \mathbf{0}_p.$$

To test these two hypotheses above, we propose two statistics as follows:

$$T_\kappa = \|\widehat{\boldsymbol{\kappa}}\|^2 = (\widehat{\boldsymbol{\kappa}})^T\widehat{\boldsymbol{\kappa}} \quad and \quad T_{\mathcal{A}} = \|\mathrm{vec}(\widehat{\mathcal{A}})\|^2 = \mathrm{vec}(\widehat{\mathcal{A}})^T\mathrm{vec}(\widehat{\mathcal{A}}).$$

Under the null hypotheses, the two statistics, T_κ and $T_{\mathcal{A}}$, are expected to be close to zero. One can reject the null hypothesis $H_{0,\kappa}$ if T_κ is large, and similarly, reject $H_{0,\mathcal{A}}$ if $T_{\mathcal{A}}$ is large. As the asymptotic normality of $\widehat{\mathcal{A}}$ under IV regularization can't be directly derived in Section 4, we propose a resampling procedure to approximate the null limiting distributions of these statistics. To accommodate the tensor data setting, we modify the bootstrap method by Cheng and Huang (2010). Specifically, we implement the bootstrap for our IV regularized minimization as follows:

- Step 1: Compute the estimators $\widehat{\boldsymbol{\kappa}}, \mathrm{vec}(\widehat{\mathcal{A}})$ through the IV regularized tensor regression by Algorithm 1.

- Step 2: For each b, genrate n $i.i.d.$ random variables $W_n^b = (W_{n1}^b, \ldots, W_{nn}^b)$ with $E(W_{ni}^b) = 1$, $E(W_{ni}^b - 1)^2 \to 1$, and $E(W_{ni}^b)^8 < \infty$ for $i = 1, \ldots, n$.

- Step 3: Solve the following W_n^b-weighted minimization and denote

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

their solutions as $\widehat{\boldsymbol{\kappa}}^b$ and $\widehat{\mathcal{A}}^b$,

$$\text{argmin} \sum_{i=1}^{n} W_{ni}^b l((\boldsymbol{\kappa}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_D), y_i) + \lambda_n \|\mathcal{A}\|_{\text{IV}},$$

The tuning parameters are the same as those selected in Step 1.

- Step 4: Repeat Step 2 and Step 3 for B times. Calculate

$$T_\kappa^b = (\widehat{\boldsymbol{\kappa}}^b - \widehat{\boldsymbol{\kappa}})^T (\widehat{\boldsymbol{\kappa}}^b - \widehat{\boldsymbol{\kappa}}) \quad \textit{and} \quad T_{\mathcal{A}}^b = \text{vec}(\widehat{\mathcal{A}}^b - \widehat{\mathcal{A}})^T \text{vec}(\widehat{\mathcal{A}}^b - \widehat{\mathcal{A}}).$$

- Compute $\widehat{p}_\kappa = b^{-1} \sum_{b=1}^{B} I(T_\kappa^b > T_\kappa)$ and $\widehat{p}_{\mathcal{A}} = b^{-1} \sum_{b=1}^{B} I(T_{\mathcal{A}}^b > T_{\mathcal{A}})$. The null hypothesis $H_{0,\kappa}$ is rejected if $\widehat{p}_\kappa$ is smaller than a prefixed significance level α, and $H_{0,\mathcal{A}}$ is rejected if $\widehat{p}_{\mathcal{A}}$ is smaller than a prefixed significance level α.

In some situations, the parameter of interest is $T'_{\mathcal{A}} = m'(\mathcal{A})$, where $m'(\cdot)$ is a function from $\mathbb{R}^p$ to $\mathbb{R}$. For example, when $\mathcal{A}$ is a $30 \times 30 \times 30$ tensor, we may want to check if the subregion in the middle is zero, i.e., $\text{vec}(\mathcal{A}_{15:20,15:20,15:20}) = \mathbf{0}$. In this case, we propose the statistic $T'_{\mathcal{A}} = \text{vec}(\mathcal{A}_{15:20,15:20,15:20})^T \text{vec}(\mathcal{A}_{15:20,15:20,15:20})$ to test the new hypothesis.

4. Theoretical Properties

The existing literature mainly focuses on the theoretical properties of tensor estimators while overlooking scalar estimators (Zhou et al., 2013; Wang et al., 2017; Li et al., 2018). Moreover, theoretical exploration of regularized generalized tensor regression models is relatively limited, especially in the complex context of IV regularization.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

In this section, we investigate the statistical properties of both tensor and scalar estimates in generalized tensor regression, with or without the use of regularization. In particular, we establish the asymptotic normality of non-regularized estimates, $\widehat{\boldsymbol{\kappa}}$ and $\widehat{\boldsymbol{A}}$, respectively (see Theorem 1). Then, we derive the error bound for coefficient estimate under general regularization (see Theorem 2). It is worth noting that the conclusion of Theorem 2 is not limited to IV regularization but is also applicable to common regularization methods, such as lasso or fusion penalties. Finally, we specifically derive the error bound for IV regularized tensor estimate $\widehat{\mathcal{A}}$ (see Theorem 3).

First, we provide a brief explanation of some notation. Let $\boldsymbol{x}_i = \text{vec}\,(\mathcal{X}_i)$. Define $\boldsymbol{A} \in \mathbb{R}^p = (\text{vec}(\boldsymbol{A}_1)^T, \text{vec}(\boldsymbol{A}_2)^T, \ldots, \text{vec}(\boldsymbol{A}_D)^T)^T$ as the true parameters, and let $f(\boldsymbol{A}) = \text{vec}\,(\mathcal{A}) = \text{vec}(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_D)$ be the function corresponding to the true $\mathcal{A}$. Then, the Jacobian matrix $\boldsymbol{J}$ of dimensions $\prod_{d=1}^D p_d \times \sum_{d=1}^D p_d R$ takes the following form:

$$\boldsymbol{J}(\boldsymbol{A}) = \partial f / \partial \boldsymbol{A} = [\boldsymbol{J}_1 \cdots \boldsymbol{J}_D]\,, \tag{4.1}$$

where $\boldsymbol{J}_d$ is the $\prod_{d=1}^D p_d \times R p_d$ Jacobian matrix with respect to $\boldsymbol{A}_d$ with the form $\boldsymbol{\Pi}_d\,[(\boldsymbol{A}_D \odot \cdots \odot \boldsymbol{A}_{d+1} \odot \boldsymbol{A}_{d-1} \odot \cdots \odot \boldsymbol{A}_1) \otimes \boldsymbol{I}_{p_d}]$, where $\boldsymbol{\Pi}_d$ is the $\prod_{d=1}^D p_d \times \prod_{d=1}^D p_d$ permutation matrix that reorders $\text{vec}(\mathcal{A}_{(d)})$ to $\text{vec}(\mathcal{A})$.

We first consider the asymptotic distribution of the estimators for tensor regression (2.3) with fixed $p_1, \ldots, p_D$ and rank R. Non-regularized tensor regression problem can be represented by $(\widehat{\boldsymbol{\kappa}}, \widehat{\boldsymbol{A}}) = \underset{\boldsymbol{\kappa}, \boldsymbol{A}}{\text{argmin}} \sum_{i=1}^n l\,\{(\boldsymbol{\kappa}, \boldsymbol{A}), y_i\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

We establish the asymptotic distribution of $\widehat{\boldsymbol{\kappa}}$ and $\widehat{\boldsymbol{A}}$ in the following theorem, with proof provided in the Supplement Material S.13.

Theorem 1. *Let* $\eta_0 = \boldsymbol{z}^T\boldsymbol{\kappa}_0 + \boldsymbol{x}^T f(\boldsymbol{A}_0)$, $\rho_2(\eta) = (d\mu/d\eta)^2/\sigma^2(\eta)$ *and define* $\boldsymbol{h}(\boldsymbol{z}) = E[\boldsymbol{z}\rho_2(\eta_0)|\mathcal{X}]/E[\rho_2(\eta_0)|\mathcal{X}]$, $\widetilde{\boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{h}(\boldsymbol{z})$ *and* $\boldsymbol{h}(\mathcal{X}) = E[\mathcal{X}\rho_2(\eta_0)|\boldsymbol{z}]/E[\rho_2(\eta_0)|\boldsymbol{z}]$, $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \mathrm{vec}(\boldsymbol{h}(\mathcal{X}))$. *Suppose Conditions C1 and C2 of the Supplement Material S.3, hold, as* $n \to \infty$, *we have*

$$\sqrt{n}(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0) \to N\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\kappa^{-1}\right), \quad \sqrt{n}(\widehat{\boldsymbol{A}} - \boldsymbol{A}_0) \to N\left(\boldsymbol{0}, \boldsymbol{\Sigma}_A^{-1}\right)$$

where $\boldsymbol{\Sigma}_\kappa = E\left[\rho_2(\eta_0)\widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^T\right]$ *and* $\boldsymbol{\Sigma}_A = \boldsymbol{J}^T(\boldsymbol{A}_0)E\left[\rho_2(\eta_0)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T\right]\boldsymbol{J}(\boldsymbol{A}_0)$.

Theorem 1 shows that the estimates $\widehat{\boldsymbol{\kappa}}$ and $\widehat{\boldsymbol{A}}$ are both $\sqrt{n}$-consistent. If $\boldsymbol{z}$ and $\mathcal{X}$ are independent, the variance terms reduce to $\boldsymbol{\Sigma}_\kappa = E\left[\rho_2(\eta_0)\boldsymbol{z}\boldsymbol{z}^T\right]$ and $\boldsymbol{\Sigma}_A = \boldsymbol{J}^T(\boldsymbol{A}_0)E\left[\rho_2(\eta_0)\boldsymbol{x}\boldsymbol{x}^T\right]\boldsymbol{J}(\boldsymbol{A}_0)$. We further characterize the interactions between the two types of covariates. In particular, the term $\boldsymbol{h}(\mathcal{X})$ is a projection of $\boldsymbol{z}$ onto $\mathcal{X}$ and $\boldsymbol{h}(\boldsymbol{z})$ is a projection of $\mathcal{X}$ onto $\boldsymbol{z}$. When there are no covariates $\boldsymbol{z}$ in the model, the asymptotic variance of $\widehat{\boldsymbol{A}}$ is the same as Theorem 2 of Zhou et al. (2013). However, their conclusion does not consider the effects of scalar covariates.

Next, we establish the theoretical properties for the recovered coefficients when a general regularization function $\mathcal{R}(\cdot)$ is added. Let $\mathcal{I} = (\boldsymbol{\kappa}, \mathcal{A})$ represent the combination of the covariate coefficients and tensor coefficients, with $\mathcal{A}_{i_1\cdots i_D} = \sum_{r=1}^R a_{1,i_1}^r \cdots a_{D,i_D}^r$ as defined by the CP decomposition (2.2). Recall that the negative log-likelihood function is defined as

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

$l(\mathcal{I}, y_i) = -[\{y_i\theta_i - \psi(\theta_i)\}/a(\phi) + c(y_i, \phi)]$. Consequently, the overall objective function combines the negative log-likelihood function and the regularization function $\mathcal{R}(\cdot)$, namely $L(\mathcal{I}|\boldsymbol{y}) = \sum_{i=1}^{n} l(\mathcal{I}, y_i) + \lambda_n \mathcal{R}(\mathcal{I})$, where λ_n is the regularization parameter. Assuming $\Theta_{\mathcal{I}}$ as the parameter space of $\mathcal{I}$, we have $\widehat{\mathcal{I}} = \text{argmin}_{\mathcal{I}\in\Theta_{\mathcal{I}}} L(\mathcal{I}|\boldsymbol{y})$. For each y_i, let $l_\Delta(\mathcal{I}, y_i) = l(\mathcal{I}, y_i) - l(\mathcal{I}_0, y_i)$ be the negative log-likelihood difference. Here, $\mathcal{I}_0 = (\boldsymbol{\kappa}_0, \mathcal{A}_0)$ corresponds to the unique true parameter. We first define $K(\mathcal{I}, \mathcal{I}_0) = n^{-1}\sum_{i=1}^{n} E\left[l_\Delta(\mathcal{I}, y_i)\right]$, which is the expected negative log-likelihood difference. Since $\mathcal{I}_0$ is the unique true parameter, we have $K(\mathcal{I}, \mathcal{I}_0) \geq 0$ for all $\mathcal{I} \in \Theta$ and $K = 0$ if and only if $\mathcal{I} = \mathcal{I}_0$. Therefore, we define the distance between $\mathcal{I}$ and $\mathcal{I}_0$ as $\rho(\mathcal{I}, \mathcal{I}_0) = K^{1/2}(\mathcal{I}, \mathcal{I}_0)$, and also define the variance of the negative log-likelihood as $V(\mathcal{I}, \mathcal{I}_0) = n^{-1}\sum_{i=1}^{n} Var\left[l_\Delta(\mathcal{I}, y_i)\right]$.

Theorem 2. *Suppose $\widehat{\mathcal{I}}$ is the regularized estimator satisfying $L(\widehat{\mathcal{I}}|\boldsymbol{y}) \leq \inf_{\mathcal{I}\in\Theta_{\mathcal{I}}} L(\mathcal{I}|\boldsymbol{y}) + \tau_n$, where $\tau_n \to 0$ and $\Theta_{\mathcal{I}} = \{(\boldsymbol{\kappa}, \mathcal{A}) : \mathcal{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_D], \|\boldsymbol{\kappa}\|_\infty \leq C_1, \max_d \|\text{vec}(\boldsymbol{A}_d)\|_\infty \leq C_2\}$ for positive constants C_1 and C_2, then*

$$P\left(\rho(\widehat{\mathcal{I}}, \mathcal{I}_0) \geq \xi_n\right) \leq 7\exp(-cn\xi_n^2),$$

where $c > 0$ is a constant and $\xi_n = \max(\varepsilon_n, \lambda_n^{1/2})$ with

$$\varepsilon_n \sim \begin{cases} \left(\frac{1}{n^{1/2}}\right)^{\frac{2\omega}{2\omega+1}} & \text{if } \omega > \frac{1}{2}, \\ \left(\frac{1}{n^{1/2}}\right)^{\omega} & \text{if } \omega \leq \frac{1}{2}, \end{cases}$$

being the best possible rate achieved when $\lambda_n \sim \varepsilon_n^2$. Here $\omega = \alpha/\gamma$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

α is the parameter associated with the degree of smoothness of l_Δ and $\gamma = \sum_{d=1}^{D} p_d R + q$ is the number of the parameters $(\boldsymbol{\kappa}, \{\boldsymbol{A}_d\}_d)$.

Theorem 2 indicates that, with an appropriately diminishing $\mathcal{R}(\cdot)$, the recovered coefficients $\widehat{\mathcal{I}}$ converge to the true one as $n \to \infty$. When l_Δ is infinitely differentiable, meaning $\omega = \infty$, the convergence rate of $\widehat{\mathcal{I}}$ reaches $n^{-1/2}$. This condition is met in cases where y follows normal, binomial, or Poisson distributions. As the smoothness of l_Δ decreases, the convergence rate of the estimator $\widehat{\mathcal{I}}$ becomes slower. Additionally, Theorem 2 does not require the penalty function $\mathcal{R}(\mathcal{I})$ to specifically be the IV as defined by (2.4). Therefore, it holds for a wide range of regularized generalized tensor regression models.

We further proceed to study the rate of convergence for IV regularized generalized tensor regression. We focus on the following estimator:

$$\widehat{\mathcal{A}} = \operatorname{argmin}_{\mathcal{A}=[\boldsymbol{A}_1,\ldots,\boldsymbol{A}_D]} \frac{1}{n} \sum_{i=1}^{n} l(\mathcal{A}, y_i) + \lambda \|\mathcal{A}\|_{\mathrm{IV}}, \tag{4.2}$$

where we use λ instead of λ_n from the previous context for distinction and simplify by ignoring the effect of covariates. To establish the error bound of $\widehat{\mathcal{A}}$ in (4.2), the following conditions are required.

(C3). The factor matrices $\mathcal{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_D]$ satiety $|S_d^r| \leq s_d^r$, where $|S_d^r|$ is the number of indices in $S_d^r = \left\{j : a_{dj+1}^r - a_{dj}^r \neq 0, 1 \leq j \leq p_d - 1\right\}$ for any $r = 1, \ldots, R$ and $d = 1, \ldots, D$.

(C4). $\mathcal{X}$ is almost surely bounded by a constant L such that $\|\operatorname{vec}(\mathcal{X})\|_\infty \leq$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

L, a.s.

(C5). The parameter space of factor matrices is $\Omega_{\mathcal{A}} = \{(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_D)$ for each $d = 1, \ldots, D$ and $r = 1, \ldots, R$, $\underline{g} \leq \|\boldsymbol{a}_d^r\|_1 \leq \bar{g}$ and $\underline{t} \leq \|\boldsymbol{a}_d^r\|_{\text{TV}} \leq \bar{t}, \}$, where $\underline{g}, \bar{g} > 0$ and $\underline{t}, \bar{t} > 0$.

(C6). Let $\Omega = \{\mathcal{H} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D} : \mathcal{H} = [\boldsymbol{A}_1 + \boldsymbol{H}_1, \boldsymbol{A}_2 + \boldsymbol{H}_2, \ldots, \boldsymbol{A}_D + \boldsymbol{H}_D]$ $- [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_D]$, where $[\boldsymbol{A}_1 + \boldsymbol{H}_1, \boldsymbol{A}_2 + \boldsymbol{H}_2, \ldots, \boldsymbol{A}_D + \boldsymbol{H}_D] \in \Omega_{\mathcal{A}}$, $\boldsymbol{h} = \text{vec}(\mathcal{H})$ and $\sum_{r=1}^{R} \sum_{d=1}^{D} (2\underline{t}^{D-1} - 1) \sum_{j \in S_d^{rc}} |h_{dj+1}^r - h_{dj}^r| \leq \sum_{r=1}^{R} \sum_{d=1}^{D} (2\bar{t}^{D-1} + 1) \sum_{j \in S_d^r} |h_{dj+1}^r - h_{dj}^r| + \sum_{r=1}^{R} \sum_{d=1}^{D} |h_{d1}^r| + \varepsilon_n / (8\bar{g}^{D-1})$ with $\varepsilon_n = 1/n\}$. Denote $\boldsymbol{\Sigma} := E(\text{vec}(\mathcal{X})^\top \text{vec}(\mathcal{X}))$, assume that there exists some constant k such that, for any $\mathcal{H} \in \Omega$,

$$\text{vec}(\mathcal{H})^\top \boldsymbol{\Sigma} \text{vec}(\mathcal{H}) \geq k \|\mathcal{H}\|_F^2 - \frac{1}{8\bar{g}^{D-1}} \varepsilon_n.$$

Condition C3 specifies the sparsity of the differences among the coefficients of the factor vectors. The s_d^r's are not required to be bounded, so the theorem holds for any value of s_d^r. Condition C4 is consistent with C1 but additionally provides the bound L for convenience. Condition C5 imposes constraints on both the range and the total variation of all factor vectors. Similar assumptions are commonly seen in regularized tensor regression problems (Lu et al., 2020; Liu et al., 2024). Note that here we constrain the lower bound of the total variation of the factor vectors, primarily due to the complex product form of the total variation in the IV regularization term. Condition C6 is similar to those in (Blazère et al., 2014; Lu et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

2020; Liu et al., 2024), but our parameter space Ω is more complex due to the consideration of the IV regularization term. C6 assumes local strong convexity of the expected loss at the minimizer, which is commonly considered in lasso or group lasso regularized GLMs (Bunea, 2008; Lounici et al., 2011; Blazère et al., 2014). By Lemma 2.1 in Bunea (2008), C6 is satisfied when the entries of $\mathbf{\Sigma}$ are bounded. The following result provides the rate of convergence for IV regularized estimate $\widehat{\mathcal{A}}$ in (4.2).

Theorem 3. *Under Conditions C3 to C6, define* $s = \max_{r,d} s_d^r$, $p = \prod_d p_d$, *and* $B = R\bar{g}^D$, *there exist constants* $C_{L,B}$, C_1, C_2, C_3, *and* C_4 *such that*

$$\lambda \geq 16KLC_{L,B}\bar{g}^{D-1}\sqrt{\frac{2\log 2p}{n}} \vee \frac{16}{3}K^2LC_{L,B}\bar{g}^{D-1}\frac{\log 2p}{n} \vee 80KL\Phi(L\zeta_n)\bar{g}^{D-1}\sqrt{\frac{2\log 2p}{n}},$$

where $K \geq 1$, $M = C_1B + \varepsilon_n$, $\zeta_n = 2M + B$, *and* $\Phi(t) = \max_{|x|\leq t}\{\psi'(x)/2\}$. *Define* $c_n = \Phi(L(M+B))$. *Then, with probability* $1 - (2 + C_2)(2p)^{-K^2/2}$, *we have*

$$\|\widehat{\mathcal{A}} - \mathcal{A}^*\|_F \leq \frac{C_3(C_4 + \sqrt{s})\sqrt{RD}\lambda}{c_n k}.$$

In particular, if $\bar{g}$ and D are bounded above and $\Phi(t)$ is bounded away from zero, the rate of convergence is $\sqrt{sR\log(2p)/n}$ when we set $\lambda \asymp \sqrt{\log(2p)/n}$. This implies consistency of the estimator when p also diverges with n as long as $\sqrt{sR\log(2p)/n} = o(1)$ as $n \to \infty$. The error bound depends on s, which represents the sparsity of the internal variation within the coefficients of the factor vectors. This implies that the more piecewise constant the true high-dimensional imaging coefficient is, the smaller the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

error bound. One may compare this to the case of Lasso penalty, where the rate depends on the sparsity of the coefficients themselves. The error bound also depends on the sample size n and the dimension p only has a logarithmic effect. The above result shows that a smaller value of s and the low-rank structure will lead to a better rate.

5. Simulation Studies

In this section, we conduct various numerical studies to evaluate the performance of our proposed method, TensorReg IV, in 2D and 3D scenarios. For comparison, we also analyze three additional methods. The first method, VoxelReg, performs generalized linear regression for each voxel. The second method, TensorReg, applies CP decomposition to $\mathcal{A}$ and estimates parameters through traditional GLMs. The third method, TensorReg Lasso, performs tensor regression with fusion lasso penalty (Li and Zhang, 2021). Considering different response distributions of y such as binomial, normal, and Poisson, we primarily present the results of simulations under the binomial distribution here, in line with our CRS data study. The 2D simulation results are provided in the Supplement Material S.7. The results for the other two distributions are compiled in the Supplement Material S.8. Estimation accuracy is assessed by $\|\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0\|_2$ and $\|\widehat{\mathcal{A}} - \mathcal{A}_0\|_F$. For prediction performance in binary tensor regression, we calculate the accuracy rate $P[\widehat{y} = y]$, sensitivity $P[\widehat{y} = 1|y = 1]$, and specificity $P[\widehat{y} = 0|y = 0]$.

In the 3D image simulation, we generate $\mathcal{X}_i$ as a $30 \times 30 \times 30$ tensor,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

incorporating four distinct shapes: One Brick, Two Bricks, Three Cross, and Pyramid, as specified in Feng et al. (2021). The ranks for these shapes are set at 1, 2, 3, and 8, respectively. The tuning parameters for regularization methods are determined through a grid search based on the criterion discussed in Section 3.1. Define $\eta_i = \boldsymbol{z}_i^T \boldsymbol{\kappa}_0 + \langle \mathcal{X}_i, \mathcal{A}_0 \rangle$, with both $\mathcal{X}_i$ and $\boldsymbol{z}_i$ being normally distributed, and we set $\boldsymbol{\kappa}_0 = (1, 1, 1, 1, 1)^T$. The binomial response is generated as $y_i \sim \text{Bernoulli}(p_i)$, with $p_i = 1/[1+\exp(-\eta_i)]$. The simulation outcomes are derived from 100 repetitions, and the sample size n varies within $\{500, 700, 1000\}$.

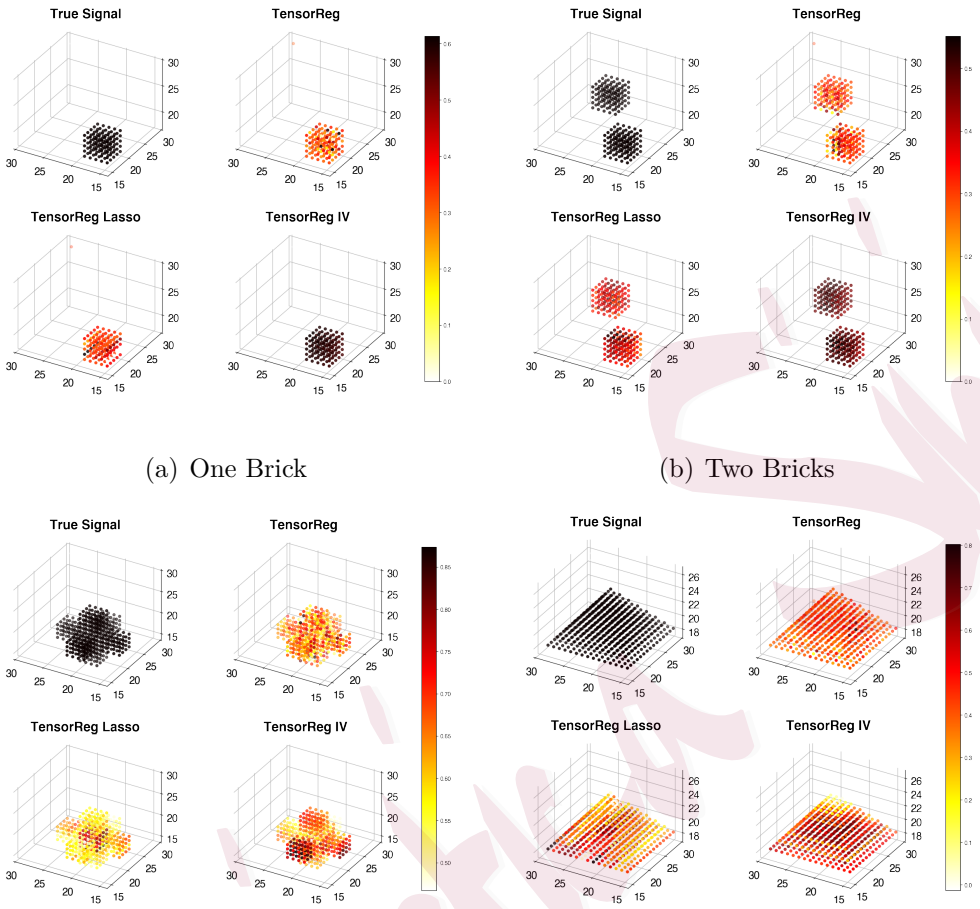
Table 1 summarizes the average RMSEs and their standard errors for $\widehat{\boldsymbol{\kappa}}$ and $\widehat{\mathcal{A}}$. The proposed method consistently outperforms the other methods in terms of $\|\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0\|_2$ and $\|\widehat{\mathcal{A}} - \mathcal{A}_0\|_F$. A notable trend is observed where the RMSEs of all methods decrease as the sample size n increases, which aligns well with our theoretical findings. The RMSEs of the VoxelReg estimates remain relatively constant across different sample sizes n, with each estimate value being near 0. This explains the anomalously small RMSE observed in the Pyramid case, highlighting it as an invalid result. To further compare the proposed method with other methods, we visualize the true signals alongside the estimated tensor coefficients of these methods in Figure 2. We omit the results of VoxelReg as it fails to reconstruct the true signal effectively. From the visualizations, it is evident that our proposed method outperforms the others in accurately recovering the true signal. Notably,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

Table 1: Mean and standard error of $\|\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0\|_2$ and $\|\widehat{\mathcal{A}} - \mathcal{A}_0\|_F$ using TensorReg IV and competing methods under different signal shapes of $\mathcal{A}_0$ and sample size n based on 100 replications.

	$\mathcal{A}_0$	n	VoxelReg	TensorReg	TensorReg Lasso	TensorReg IV
$\|\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0\|_2$	One brick	500	2.199 (0.006)	1.629(0.904)	1.514(0.825)	**1.172(0.525)**
		700	2.185(0.007)	1.512(0.621)	1.368(0.883)	**0.641(0.477)**
		1000	2.162(0.008)	0.738(0.512)	0.811(0.514)	**0.355(0.174)**
	Two bricks	500	2.210(0.006)	1.273(0.428)	1.533(0.290)	**1.354(0.411)**
		700	2.198(0.007)	1.481(0.232)	1.169(0.437)	**0.883(0.489)**
		1000	2.181(0.009)	1.293(0.437)	0.678(0.374)	**0.582(0.368)**
	Three cross	500	2.215(0.007)	2.186(0.643)	1.551(0.519)	**1.331(0.375)**
		700	2.208(0.008)	1.524(0.594)	1.317(0.353)	**1.117(0.380)**
		1000	2.195(0.010)	1.444(0.439)	1.065(0.281)	**0.820(0.335)**
	Pyramid	500	2.218(0.006)	1.928(0.194)	1.696(0.249)	**1.505(0.330)**
		700	2.213(0.008)	1.730(0.315)	1.606(0.265)	**1.294(0.333)**
		1000	2.203(0.010)	1.490(0.360)	1.485(0.249)	**1.105(0.300)**
$\|\widehat{\mathcal{A}} - \mathcal{A}_0\|_F$	One brick	500	6.690(0.005)	8.952(5.665)	7.714(2.859)	**4.741(2.655)**
		700	6.681(0.006)	7.023(2.506)	6.633(3.667)	**2.162(2.289)**
		1000	6.672(0.006)	3.658(1.991)	3.845(2.008)	**0.985(0.598)**
	Two bricks	500	9.436(0.005)	18.381(21.921)	9.706(0.334)	**8.193(2.221)**
		700	9.416(0.005)	9.896(0.379)	8.087(1.301)	**4.841(2.698)**
		1000	9.387(0.006)	8.405(1.595)	5.454(1.598)	**2.918(1.957)**
	Three cross	500	12.631(0.004)	24.172(6.806)	19.532(4.837)	**11.353(1.665)**
		700	12.604(0.005)	29.909(17.731)	14.265(4.746)	**9.196(1.684)**
		1000	12.562(0.006)	18.028(18.738)	9.452(1.326)	**6.438(2.281)**
	Pyramid	500	**15.567(0.004)**	16.449(3.68)	19.209(1.306)	15.965(6.704)
		700	15.536(0.005)	17.214(4.122)	17.670(3.202)	**12.056(1.490)**
		1000	15.489 (0.006)	14.681 (3.726)	13.587(2.311)	**10.361(0.875)**

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281



(a) One Brick (b) Two Bricks

(c) Three Cross (d) Pyramid

Figure 2: True signals and estimates of $\widehat{\mathcal{A}}$. The sample size is 1000.

TensorReg and TensorReg Lasso struggle to reconstruct the true signal, even in the relatively simple One Brick scenario.

Table 2 details the prediction performance of various methods. A similar pattern to the estimation accuracy results is evident: TensorReg IV consistently outperforms the other methods, achieving higher accuracy, sensitivity, and specificity across the board. In terms of computation time, our method takes longer than TensorReg but is comparable to TensorReg Lasso.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

For instance, in the One Brick scenario with $n = 1000$, the average computation times are 95 seconds for TensorReg, 300 seconds for TensorReg Lasso, and 284 seconds for IV regularization. Based on the valuable suggestion from the AE, we also compare the performance of CNNs on the aforementioned classification problem, with results presented in Table 2. The training and tuning steps for the CNNs are detailed in the Supplement Material S.5. It can be observed that the performance of CNNs is generally suboptimal. Due to the data generation process, our images are generated from a normal distribution, making it challenging for CNNs to capture label-related signals, resulting in suboptimal performance. The sensitivity and specificity results are provided in the Supplement Material S.6.

6. The CRS Application

The etiology of CRS is complex, and it is a common condition worldwide with a prevalence of about 5%-16%. CRS significantly impairs quality of life and poses a substantial economic burden (Hastan et al., 2011; Fokkens et al., 2020; Zou et al., 2024). Diagnosis primarily relies on nasal endoscopy and imaging assessment, in addition to clinical symptoms such as nasal congestion and blockage. CT examination can display all groups of sinus cavities, mucosa, and bony structures without the restriction of field of view, making it essential for preoperative evaluation. Combining CT image data with patient-related covariates significantly enriches diagnostic information and improves accuracy.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

Table 2: Mean and standard error of prediction results using TensorReg IV and competing methods under different signal shapes of $\mathcal{A}_0$ and sample size n based on 100 replications.

	$\mathcal{A}_0$	n	VoxelReg	TensorReg	TensorReg Lasso	CNN	TensorReg IV
$P[\widehat{y}=y]$	One brick	500	0.514(0.034)	0.579(0.119)	0.575(0.108)	0.538 (0.053)	**0.662(0.170)**
		700	0.525(0.027)	0.683(0.170)	0.745(0.168)	0.566 (0.061)	**0.814(0.142)**
		1000	0.532(0.025)	0.787(0.154)	0.779(0.155)	0.611 (0.065)	**0.879(0.039)**
	Two bricks	500	0.519(0.035)	0.512(0.034)	0.553(0.061)	0.540 (0.049)	**0.629(0.124)**
		700	0.523(0.032)	0.538(0.042)	0.699(0.118)	0.559 (0.057)	**0.804(0.130)**
		1000	0.526(0.022)	0.646(0.127)	0.845(0.088)	0.599 (0.054)	**0.864(0.087)**
	Three cross	500	0.517(0.036)	0.524(0.042)	0.595(0.060)	0.562 (0.067)	**0.659(0.096)**
		700	0.526(0.030)	0.533(0.055)	0.678(0.044)	0.615 (0.077)	**0.734(0.072)**
		1000	0.532(0.021)	0.601(0.091)	0.737(0.024)	0.654 (0.072)	**0.826(0.077)**
	Pyramid	500	0.520(0.031)	0.533(0.082)	0.600(0.038)	0.568 (0.077)	**0.701(0.048)**
		700	0.521(0.029)	0.583(0.085)	0.635(0.043)	0.629 (0.083)	**0.750(0.036)**
		1000	0.532(0.021)	0.654(0.092)	0.703(0.043)	0.660 (0.090)	**0.777(0.036)**

We apply the proposed method to a real dataset of CRS patients. The dataset, comprising clinical and imaging data, was collected from patients who visited Nanjing Tongren Hospital for CT examinations of paranasal sinuses between January 2018 and December 2021. A retrospective analysis of these patients and a healthy control group was conducted. All subjects underwent Multi-Slice CT (MSCT) as part of their assessment. CRS diagnoses followed the European sinusitis guidelines. For image acquisition, CT volumetric images of the patients' paranasal sinuses were obtained using either a 64-layer or 256-layer spiral CT scanner (Philips Medical Systems,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

The Netherlands). These images were captured with a layer thickness and spacing of 0.625 mm, and a size of 512×512. The scanning baseline was aligned parallel to the infraorbital line, covering an area from the top of the frontal sinus down to the inferior edge of the maxillary odontoid process.

Due to varying sizes of CT images among patients and to improve computation efficiency, we cropped the CT images for each patient, removing irrelevant areas, and resized them to $100 \times 256 \times 256$. The dataset includes 479 cases: 270 normal controls and 209 CRS subjects. Of these, 274 were males with an average age of 35.5 years (SD 17.2 years) and 205 were females with an average age of 40.7 years (SD 17.3 years). We encoded the binary disease state as 0 for healthy controls (HC) and 1 for CRS. The image predictor $\mathcal{X}_i$ is a 3D CT image of the sinus, and the covariate vector $\boldsymbol{z}_i$ includes gender (female=0, male=1) and age (ranging from 3 to 81 years). Given $(\boldsymbol{z}_i, \mathcal{X}_i)$, y_i is assumed to follow a Bernoulli distribution with probability p_i, where $\log[p_i/(1-p_i)] = \boldsymbol{z}_i^T \boldsymbol{\kappa} + \langle \mathcal{X}_i, \mathcal{A} \rangle$. We employed Algorithm 1 to estimate the unknown parameters.

Figure 3 presents the horizontal, coronal, and sub-regional sagittal sections as identified by three different methods. Specifically, the first row in Figure 3 displays three sections of an original CT image sample. The sixth color row shows the estimated coefficients obtained by our proposed method, and the seventh row highlights the signal regions within the top 10% of the magnitude. Notably, in the bottom left panel of Figure 3, the ef-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

fects around pixels $(50, 100)$ and $(50, 150)$ appear to be effectively captured by IV estimation. Similar observations are made in the coronal and sub-regional sagittal planes, where the effects at corresponding locations are also well estimated. In contrast, the signal distribution in the estimated coefficients obtained by TensorReg and TensorReg Lasso is more dispersed, making it challenging to intuitively discern significant signal regions.

Analyzing the bottom panel of Figure 3, we can locate the subregion with the strongest signal, identified by TensorReg IV as $\widehat{\mathcal{A}}_{20:30,155:170,140:150}$. We test the null hypothesis $H_{0,\mathcal{A}'}: \text{vec}(\widehat{\mathcal{A}}_{20:30,155:170,140:150}) = \mathbf{0}$ and propose the corresponding statistic $T_{\mathcal{A}'} = \text{vec}(\widehat{\mathcal{A}}_{20:30,155:170,140:150})^T \text{vec}(\widehat{\mathcal{A}}_{20:30,155:170,140:150})$. We calculated the p-value of this subregion using the empirical bootstrap procedure proposed in Section 3.2. The p-value for this subregion is 0.020, indicating that the identified subregion is significant on the incidence of CRS. Regarding the covariates, we obtained $\widehat{\boldsymbol{\kappa}}$ of the coefficients corresponding to age and sex as -0.026 and -1.262, respectively, with the corresponding p-values being 0.008 and 0.739, which indicates that sex does not have a significant impact on the incidence of CRS.

It is noteworthy that the subregion with the strongest signal identified by our method is located in the sinus cavity of the maxillary sinus, a region widely recognized for its association with the pathogenesis of CRS. Conventional CT scans for CRS typically reveal a mucosal thickness of ≥ 3 mm in the maxillary sinus cavity, often accompanied by soft tissue and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

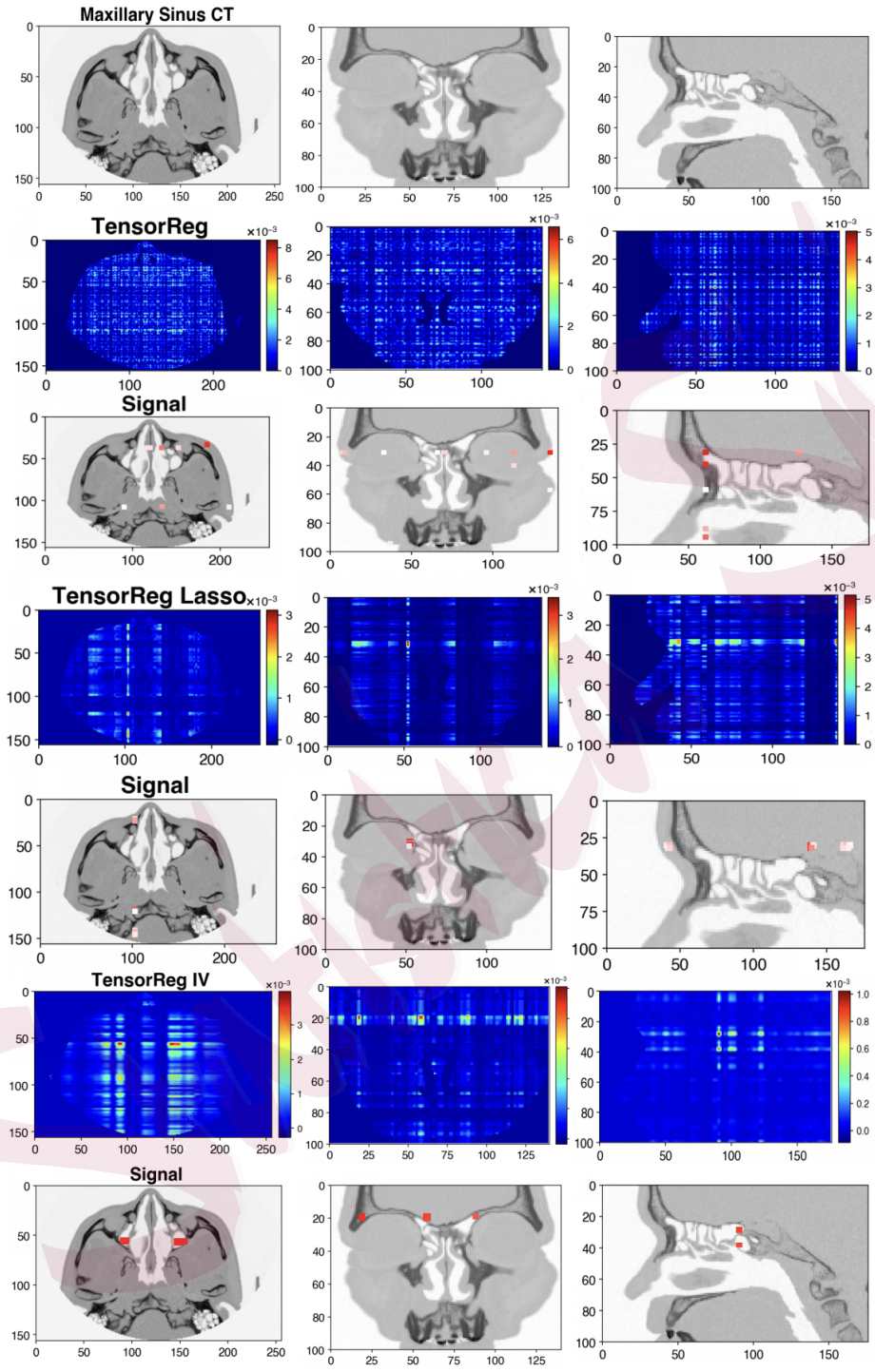


Figure 3: Tensor coefficients estimated and identified signals for CRS data using different methods. Three columns are the horizontal section, the coronal section, and sagittal section, respectively. Row 1 is a sample of sinus CT, rows 2-3, rows 4-5, rows 6-7 correspond to TensorReg, TensorReg Lasso and TensorReg IV, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

air-fluid surfaces (Capelli and Gatti, 2016; Fokkens et al., 2020). Classical diagnostic criteria assign scores of 0, 1, or 2 for the absence, presence, or complete filling of inflammation in each sinus cavity, respectively. Additionally, studies have shown that the maxillary sinus cavity volume and mean bone wall thickness in CRS patients significantly differ from those in the control group (Kim et al., 2008; Cho et al., 2010; Deeb et al., 2011). The scalar estimates also provide insightful biological meanings. Firstly, the negative estimate for age indicates a lower prevalence of CRS in older groups compared to younger ones, aligning with the results from a CRS symptom questionnaire administered to 24,000 primary care patients (Mahdavinia and Grammer III, 2013; Hirsch et al., 2017). Secondly, regarding the negative gender coefficient, although it is not statistically significant, we still observed that its negative nature aligns with some literature on the relationship between CRS and gender (Shashy et al., 2004; Mahdavinia and Grammer III, 2013). This suggests a potential increased susceptibility of women to CRS, a hypothesis that warrants further investigation.

Finally, we evaluated the prediction accuracy of the proposed method and the competing methods for the CRS dataset. We randomly divided the CRS dataset into a training set with $n_1 = 400$ and a test set with $n_2 = 79$, calculating the classification accuracy for the test set. The prediction results are summarized in Table 3. Evidently, TensorReg IV achieves superior performance in terms of accuracy, sensitivity, and specificity. Furthermore,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

Table 3: Mean and standard error of prediction performance using the IV and competing methods.

Prediction	TensorReg	TensorReg Lasso	TensorReg IV
Accuracy	87.9%(0.9%)	89.5%(1.9%)	93.2%(1.1%)
Sensitivity	90.2%(0.7%)	92.3%(4.4%)	93.7%(4.2%)
Specificity	86.2%(1.8%)	86.7%(2.9%)	92.8%(4.6%)

when diagnosing CRS using only gender and age, the accuracy, sensitivity, and specificity were 50.6%, 43.6%, and 57.5%, respectively, indicating that the inclusion of high-order images substantially improves predictive power.

7. Discussion

In this article, we introduce a novel IV regularized tensor regression framework that incorporates a low-rank and piecewise constant structure. This framework is robust to distributional assumptions and enhances the interpretability of the model. We delve into the details of how IV effectively assigns varied regularization strengths to each $\{\|\boldsymbol{a}_d^r\|_{\mathrm{TV}}\}_{r,d}$. We investigate the theoretical properties of tensor estimate under IV regularization. Extensive numerical studies have been conducted to validate the effectiveness of the proposed IV regularized regression and compare it with other methods. We also applied this method to analyze a real CRS dataset, successfully identifying the most active regions associated with CRS. This work lays the groundwork for further research into the etiology and imaging interac-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

tions of CRS and other sinus diseases, a task complicated by the complexity of the sinuses and the irregular spatial structure and dimensionality of the imaging data. There remain many areas ripe for further research. The IV is based on the CP decomposition, prompting the consideration of other common tensor decomposition methods for constructing new IV, such as Tucker decomposition (Kolda and Bader, 2009) and Tensor-Train decomposition (Oseledets, 2011).

8. Supplementary Material

In the supplementary material, we present the complete algorithm for estimating the parameters and provide additional numerical results. In addition, we give some useful lemmas and proofs of theorems.

9. Acknowledgment

The authors thank the reviewers, associate editor, and co-editor for their helpful suggestions and comments. In addition, Li's research is partially supported by the National Science Foundation of China, Grant 12571304, the Shanghai Pujiang Programme (No. 24PJC030), and the Program for Innovative Research Team of Shanghai University of Finance and Economics. Bai's research is supported by the Program for Innovative Research Team of Shanghai University of Finance and Economics and the Shanghai Research Center for Data Science and Decision Technology.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

References

Bi, X., X. Tang, Y. Yuan, Y. Zhang, and A. Qu (2021). Tensors in statistics. *Annual review of statistics and its application 8*(1), 345–368.

Blazère, M., J.-M. Loubes, and F. Gamboa (2014). Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory 60*(4), 2303–2318.

Boyd, S., N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning 3*(1), 1–122.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research 33*(2), 261–304.

Capelli, M. and P. Gatti (2016). Radiological study of maxillary sinus using cbct: relationship between mucosal thickening and common anatomic variants in chronic rhinosinusitis. *Journal of Clinical and Diagnostic research: JCDR 10*(11), MC07.

Cheng, G. and J. Z. Huang (2010). Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics 38*(5), 2884–2915.

Cho, S. H., T. H. Kim, K. R. Kim, J.-M. Lee, D.-K. Lee, J.-H. Kim, J.-J. Im, C.-J. Park, and K.-G. Hwang (2010). Factors for maxillary sinus volume and craniofacial anatomical features in adults with chronic rhinosinusitis. *Archives of Otolaryngology–Head & Neck Surgery 136*(6), 610–615.

Deeb, R., P. N. Malani, B. Gil, K. Jafari-Khouzani, H. Soltanian-Zadeh, S. Patel, and M. A. Zacharek (2011). Three-dimensional volumetric measurements and analysis of the maxillary sinus. *American journal of Rhinology & Allergy 25*(3), 152–156.

Feng, L., X. Bi, and H. Zhang (2021). Brain regions identified as being associated with verbal reasoning through the use of imaging regression via internal variation. *Journal of the American Statistical Association 116*(533), 144–158.

Fokkens, W. J., V. J. Lund, C. Hopkins, P. W. Hellings, R. Kern, S. Reitsma, S. Toppila-Salmi, M. Bernal-Sprekelsen, J. Mullol, I. Alobid, et al. (2020). European position paper on rhinosinusitis and nasal polyps 2020. *Rhinology 58*(Suppl S29), I–+.

Hastan, D., W. Fokkens, C. Bachert, R. Newson, J. Bislimovska, A. Bockelbrink, P. Bousquet, G. Brozek, A. Bruno, S. Dahlén, et al. (2011). Chronic rhinosinusitis in europe–an underestimated disease. a ga2len study. *Allergy 66*(9), 1216–1223.

Hinshaw, S. P. and R. M. Scheffler (2014). *The ADHD explosion: Myths, medication, money, and today's push for performance*. Oxford University Press.

Hirsch, A. G., W. F. Stewart, A. S. Sundaresan, A. J. Young, T. L. Kennedy, J. Scott Greene, W. Feng, B. K. Tan, R. P. Schleimer, R. C. Kern, et al. (2017). Nasal and sinus symptoms and chronic rhinosinusitis in a population-based sample. *Allergy 72*(2), 274–281.

Jack, C. R., D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski (2010). Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology 9*(1), 119–128.

Kim, H. Y., M.-B. Kim, H.-J. Dhong, Y. G. Jung, J.-Y. Min, S.-K. Chung, H. J. Lee, S. C. Chung, and N. G. Ryu (2008). Changes of maxillary sinus volume and bony thickness of the paranasal sinuses in longstanding pediatric chronic rhinosinusitis. *International journal of Pediatric Otorhinolaryngology 72*(1), 103–108.

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

Li, C. and H. Zhang (2021). Tensor quantile regression with application to association between

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

neuroimages and human intelligence. *The Annals of Applied Statistics 15*(3), 1455 – 1477.

Li, X., D. Xu, H. Zhou, and L. Li (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences 10*(3), 520–545.

Liu, Y., J. Liu, Z. Long, C. Zhu, Y. Liu, J. Liu, Z. Long, and C. Zhu (2022). *Tensor regression*. Springer.

Liu, Z., C. Y. Lee, and H. Zhang (2024). Tensor quantile regression with low-rank tensor train estimation. *The Annals of Applied Statistics 18*(2), 1294–1318.

Lounici, K., M. Pontil, S. Van De Geer, and A. B. Tsybakov (2011). Oracle inequalities and optimal inference under group sparsity.

Lu, W., Z. Zhu, and H. Lian (2020). High-dimensional quantile tensor regression. *Journal of Machine Learning Research 21*(250), 1–31.

Luo, Y. and A. R. Zhang (2024). Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap and their interplay. *The Annals of Statistics 52*(6), 2583–2612.

Mahdavinia, M. and L. C. Grammer III (2013). Chronic rhinosinusitis and age: is the pathogenesis different? *Expert review of Anti-Infective Therapy 11*(10), 1029–1040.

Michel, V., A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion (2011). Total variation regularization for fmri-based prediction of behavior. *IEEE transactions on medical imaging 30*(7), 1328–1340.

Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing 33*(5), 2295–2317.

Raskutti, G., M. Yuan, and H. Chen (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics 47*(3), 1554–1584.

Shashy, R. G., E. J. Moore, and A. Weaver (2004). Prevalence of the chronic sinusitis diagnosis in olmsted county, minnesota. *Archives of Otolaryngology–Head & Neck Surgery 130*(3), 320–323.

Spencer, D., R. Guhaniyogi, and R. Prado (2019). Bayesian mixed effect sparse tensor response regression model with joint estimation of activation and connectivity. *arXiv preprint arXiv:1904.00148*.

Su, J., W. Byeon, J. Kossaifi, F. Huang, J. Kautz, and A. Anandkumar (2020). Convolutional tensor-train lstm for spatio-temporal learning. *Advances in Neural Information Processing Systems 33*, 13714–13726.

Wang, X., H. Zhu, and A. D. N. Initiative (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association 112*(519), 1156–1168.

Wu, S. and L. Feng (2023). Sparse kronecker product decomposition: a general framework of signal region detection in image regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(3), 783–809.

Yan, H., K. Paynabar, and M. Pacella (2019). Structured point cloud data analysis via regularized tensor regression for process modeling and optimization. *Technometrics 61*(3), 385–395.

Yang, Y. and T. M. Hospedales (2017). Deep multi-task representation learning: A tensor factorisation approach. In *International Conference on Learning Representations*.

Yu, R., G. Li, and Y. Liu (2018). Tensor regression meets gaussian processes. In *International conference on artificial intelligence and statistics*, pp. 482–490. PMLR.

Zhang, J., Y. Cai, Z. Wang, and B. Wang (2019). Sparse and low-rank high-order tensor regression via parallel proximal method. *arXiv preprint arXiv:1911.12965*.

Zhang, T., Y. Fu, and C. Li (2021). Hyperspectral image denoising with realistic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2248–2257.

Zhou, H., L. Li, and H. Zhu (2013). Tensor regression with applications in neuroimaging data

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0281

analysis. *Journal of the American Statistical Association 108*(502), 540–552.

Zhou, I. Y., Y.-X. Liang, R. W. Chan, P. P. Gao, J. S. Cheng, Y. Hu, K.-F. So, and E. X. Wu (2014). Brain resting-state functional mri connectivity: morphological foundation and plasticity. *Neuroimage 84*, 1–10.

Zou, C., H. Ji, J. Cui, B. Qian, Y.-C. Chen, Q. Zhang, S. He, Y. Sui, Y. Bai, Y. Zhong, et al. (2024). Preliminary study on ai-assisted diagnosis of bone remodeling in chronic maxillary sinusitis. *BMC Medical Imaging 24*(1), 140.