



## Brain Regions Identified as Being Associated With Verbal Reasoning Through the Use of Imaging Regression via Internal Variation

Long Feng, Xuan Bi & Heping Zhang

To cite this article: Long Feng, Xuan Bi & Heping 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, DOI: [10.1080/01621459.2020.1766468](https://doi.org/10.1080/01621459.2020.1766468)

To link to this article: <https://doi.org/10.1080/01621459.2020.1766468>



View supplementary material [↗](#)



Published online: 08 Jun 2020.



Submit your article to this journal [↗](#)



Article views: 910



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



# Brain Regions Identified as Being Associated With Verbal Reasoning Through the Use of Imaging Regression via Internal Variation

Long Feng<sup>\*a</sup> , Xuan Bi<sup>\*b</sup> , and Heping Zhang<sup>a</sup>

<sup>a</sup>Department of Biostatistics, Yale University, New Haven, CT; <sup>b</sup>Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, MN

## ABSTRACT

Brain-imaging data have been increasingly used to understand intellectual disabilities. Despite significant progress in biomedical research, the mechanisms for most of the intellectual disabilities remain unknown. Finding the underlying neurological mechanisms has proved difficult, especially in children due to the rapid development of their brains. We investigate verbal reasoning, which is a reliable measure of an individual's general intellectual abilities, and develop a class of high-order imaging regression models to identify brain subregions which might be associated with this specific intellectual ability. A key novelty of our method is to take advantage of spatial brain structures, and specifically the piecewise smooth nature of most imaging coefficients in the form of high-order tensors. Our approach provides an effective and urgently needed method for identifying brain subregions potentially underlying certain intellectual disabilities. The idea behind our approach is a carefully constructed concept called internal variation (IV). The IV employs tensor decomposition and provides a computationally feasible substitution for total variation, which has been considered suitable to deal with similar problems but may not be scalable to high-order tensor regression. Before applying our method to analyze the real data, we conduct comprehensive simulation studies to demonstrate the validity of our method in imaging signal identification. Next, we present our results from the analysis of a dataset based on the Philadelphia Neurodevelopmental Cohort for which we preprocessed the data including reorienting, bias-field correcting, extracting, normalizing, and registering the magnetic resonance images from 978 individuals. Our analysis identified a subregion across the cingulate cortex and the corpus callosum as being associated with individuals' verbal reasoning ability, which, to the best of our knowledge, is a novel region that has not been reported in the literature. This finding is useful in further investigation of functional mechanisms for verbal reasoning. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

## ARTICLE HISTORY

Received August 2019  
Accepted May 2020

## KEYWORDS

Brain imaging; Internal variation; Piecewise smoothness; Tensor regression; Verbal reasoning

## 1. Introduction



### 1.1. Background

Intellectual disability, formerly known as mental retardation, affects about 2%–3% of the general population (Daily, Ardinger, and Holmes 2000). It is a condition in the brain characterized by impaired intellectual, social and adaptive functioning with onset in childhood. Children with intellectual disability usually have an intelligence quotient (IQ) score under 70 on a standardized measure, and may have certain limitations in mental functioning such that they may have difficulties in communicating, understanding, problem-solving and learning self-care skills.


While genetic disorder and iodine deficiency are globally common causes, studies have found that more than 60% of intellectual disability are idiopathic intellectual disability (Vos et al. 2015), which has unknown cause and spontaneous origin. In fact, finding the neurological mechanisms underlying this trait has been proved difficult, especially given the rapid


development of brain in childhood (Shaw et al. 2006). Many valuable attempts and contributions have been made to investigate the effects of brain structure and functions on intellectual disability. For example, Shaw et al. (2006) demonstrated the relation between the level of intelligence and the change in the thickness of the cerebral cortex, and Baglio et al. (2014) investigated the effect of brain morphometry on IQ in children with borderline intellectual functioning. Meanwhile, many works investigated a specific learning disability or an adaptive limitation, rather than the general intelligence level. For instance, Gur et al. (2000) studied the sex difference in regional brain activation in verbal-reasoning and spatial tasks. Bathelt et al. (2016) investigated individual's structural brain differences in a single gene disorder, which is associated with speech and language impairments.

Studies of brain structures have benefited tremendously from the advance of neuroimaging technologies, such as magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI, Good et al. 2001; Martin and Chao 2001).

**CONTACT** Heping Zhang  [heping.zhang@yale.edu](mailto:heping.zhang@yale.edu)  Department of Biostatistics, Yale University School of Public Health, New Haven, CT 06520.

\*The first two authors contributed equally.

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

 These materials were reviewed for reproducibility.

© 2020 American Statistical Association

Some common practices of analyzing brain images especially the cortical regions include, but are not limited to, voxel-based methods (e.g., Shen et al. 2010; Zhou et al. 2014) and surface-based methods. The latter may provide accurate local mapping of the cortical morphology. For example, existing works have demonstrated the effects of cortical thickness (Burgaleta et al. 2014), cortical surface area (Schnack et al. 2015), cortical volume (Bajaj et al. 2018), and subcortical gray matter and white matter (Bathelt et al. 2016) on intellectual ability or related specific abilities such as perceptual reasoning and verbal comprehension.

Detecting the association between intellectual disability and brain structures using brain imaging also poses great challenges, due to the complex, high-order, high-resolution nature of imaging data. Most commonly used practices usually use imaging voxels as independent covariates, calculate intensity difference between case and control groups, or consider region or volume based variables via parcellation or segmentation of images. The implementation of these methods is straightforward and may provide valuable inferences regarding the interested brain regions. However, neuroimaging data may contain rich spatial and structural information about the brain, and the latent associations (or synergies) across or at the boundary of different brain anatomical regions might be omitted due to the high-level summarization (e.g., aggregation or averaging) imposed by the existing methods. An effective statistical method which can exploit the raw yet rich information from the neuroimaging data is urgently needed. And this is the motivation of this article.

## 1.2. Philadelphia Neurodevelopmental Cohort (PNC)

The PNC is a large-scale research initiative. One of its goals is to understand the impact of brain development, genetics, and environment on the cognitive abilities of children and adolescents and their vulnerability to mental disorders (Satterthwaite et al. 2014). The research data were collected from over 9500 individuals from the greater Philadelphia area with age ranging from 8 to 21 years and medical conditions ranging from no medical condition to having chronic or potentially life-threatening diseases. The participants were genotyped during their clinical visit upon consent, and were selected at random after a population stratification using age, ethnicity, and sex.

All participants were screened for psychiatric disorders and were assessed through a structural interview and a Computerized Neurocognitive Battery to quantify their performance in multiple cognitive domains, including complex and social cognition, episodic memory, executive control, and sensorimotor speed. Among all PNC participants, 1445 individuals who were interested in participating and met eligibility criteria also received additional neuroimaging scans. The scans include structural and functional MRI, as well as diffusion tensor imaging. In particular, all neuroimages were acquired at the same site, on a single scanner, and without any software updates, such that the differences among neuroimages caused by the scanner are reduced to the minimum. More details about the PNC initiative are provided in Gur et al. (2010, 2012), and at <https://www.med.upenn.edu/bbl/philadelphianeurodevelopmentalcohort.html>.

## 1.3. Existing Works on Statistical Analysis of Neuroimaging Data

Statistics plays an important role in many stages of neuroimaging research, ranging from acquiring raw neuroimaging data, locating brain activity, to predicting psychological, psychiatric or cognitive states (Lindquist 2008). The topic, statistical analysis of neuroimaging data, has drawn a great deal of attention in the statistical literature over the past decades (e.g., Zhou, Li, and Zhu 2013). From a statistical point of view, brain imaging data are typically stored in multidimensional arrays, or tensors, with each element representing a voxel. The value of each element, usually nonnegative, shows the signal (brightness) of the image. A straightforward approach to analyzing such imaging data is through voxelwise regression (e.g., Shen et al. 2010; Zhou et al. 2014), where an interested phenotype is regressed against each voxel of the brain image. One obvious drawback of this approach is that structural information and connectivity among brain regions are not considered. In addition, with image resolution getting increasingly high, voxelwise regression suffers from a severe computational burden. For example, the number of voxels in an MRI image could possibly range from  $64^3 = 262,144$  to  $256^3 = 16,777,216$ . In an fMRI image, the number of voxels could be even larger.

Zhou, Li, and Zhu (2013) proposed an innovative imaging regression framework, named tensor regression. They impose a canonical polyadic decomposition (CPD) on the tensor coefficients. By using a CPD, their model not only reduces the dimensionality of the tensor coefficients tremendously and hence eases the computation, but also retains the structural information of the brain. While their tensor regression approach is appealing, it does not address one important characteristic of imaging data—piecewise smoothness. Piecewise smoothness describes the fact that voxels in close proximity are more likely to have close coefficients, even if they are at or across the boundary of brain anatomical regions. In our context, this characteristic is relevant to the capability of identifying subregions which may not necessarily belong to a specific anatomical region, or synergies across multiple brain regions.

Total variation (TV) has been one of the most commonly used approaches for image denoising and recovery since the pioneering work of Rudin, Osher, and Fatemi (1992) and Rudin and Osher (1994). More recently, TV is also used to control the smoothness of image coefficients in regression models (Michel et al. 2011; Wang, Zhu, and Alzheimer's Disease Neuroimaging Initiative 2017). It is known that TV-based approaches are powerful and efficient for two-dimensional image regression. However, it might be difficult to apply them to third or higher order tensor images due to both computational and analytical complexities.

## 1.4. Overview

In this article, we investigate one specific intellectual ability—verbal reasoning, which is known as a reliable measure of individuals' general intellectual abilities (Gur et al. 2010). Childhood difficulty in verbal reasoning can be considered as a special type of language disorder, which is a common feature of neurodevelopmental disorder.

We propose a general yet innovative method, named shrinkage via internal variation (SHIV), which can identify and locate brain regions that are directly associated with verbal reasoning. The proposed method is completely different from the surface-based and voxel-based methods which are commonly used by practitioners. It adopts the imaging regression framework yet accounts for the piecewise smoothness.

Specifically, to overcome the difficulty of allowing both high-order tensor images and piecewise smooth coefficients, we first propose an innovative quantity named internal variation (IV), which aims to substitute TV for high-order imaging data. Unlike TV that performs the component-wise calculation, IV makes use of the CPD of tensor and intrinsically preserves the structural information of image. The SHIV is an IV penalized minimization in the imaging regression framework. Compared to TV-based approaches, SHIV can not only produce smooth tensor coefficients, but is also computationally more feasible for high-order tensor image as it can be solved easily by a sequence of fused lasso (Tibshirani and Taylor 2011) problems.

Numerically, the proposed SHIV approach is tested carefully on the simulated data and the PNC data, where the primary goal is to locate brain regions which may be associated with each individual's verbal reasoning ability. Brain MRI images from 978 PNC participants are selected and preprocessed, which include reorientation, bias-field correction, brain extraction, intensity normalization, and linear registration. The proposed SHIV method clearly identifies a subregion across the cingulate cortex and the corpus callosum as being associated with individual's verbal reasoning ability, which, to the best of our knowledge, is a novel region that has not been reported in the literature. This finding is useful in further investigation of functional mechanisms for verbal reasoning.

The remainder of the article is organized as follows. In Section 2, we introduce the general method, including IV, SHIV, and the computation. Section 3 provides simulation studies. In Section 4, we analyze the PNC dataset along with a PNC data based simulation. We conclude with a few remarks in Section 5.

## 2. Method

### 2.1. Scalar-on-Image Regression and Total Variation

A multidimensional array  $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times \dots \times p_D}$  is called a  $D$ th order tensor, with each dimension of the array named as a mode. Here  $p_1, p_2, \dots, p_D$  represent the size (number of voxels) of the tensor image along each mode. For instance, an MRI scan produces a third-order tensor, and an fMRI scan or multimodality imaging produces an even higher-order ( $D > 3$ ) tensor. Given an observed  $D$ -dimensional tensor-image covariate  $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \dots \times p_D}$ , a  $q$ -dimensional vector-valued auxiliary covariate  $\mathbf{z} \in \mathbb{R}^q$ , and a scalar-valued continuous response  $y$ , a linear imaging regression can be written as

$$y = \mathbf{z}^\top \boldsymbol{\theta}^* + \langle \mathcal{X}, \mathcal{B}^* \rangle + \epsilon, \quad (1)$$

where  $\epsilon$  is an error term,  $\boldsymbol{\theta}^* \in \mathbb{R}^q$  and  $\mathcal{B}^* \in \mathbb{R}^{p_1 \times p_2 \times \dots \times p_D}$  are the vector and tensor coefficients need to be estimated. Here  $\langle \cdot, \cdot \rangle$  denotes element-wise inner product. For example, in a simple

three-dimensional image with  $p_1 = p_2 = p_3 = 2$  and

$$\mathcal{X}_{::,1} = \begin{pmatrix} x_{111} & x_{121} \\ x_{211} & x_{221} \end{pmatrix}, \quad \mathcal{X}_{::,2} = \begin{pmatrix} x_{112} & x_{122} \\ x_{212} & x_{222} \end{pmatrix},$$

$$\mathcal{B}_{::,1}^* = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathcal{B}_{::,2}^* = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

with element-wise inner product, Equation (1) becomes

$$y = \mathbf{z}^\top \boldsymbol{\theta}^* + 1 \cdot x_{121} + 1 \cdot x_{211} + 2 \cdot x_{221} + 1 \cdot x_{222} + \epsilon.$$

In practice, the values in  $\boldsymbol{\theta}^*$  and  $\mathcal{B}^*$  are unknown.

This article focuses on finding the connections between verbal reasoning and brain structure. The outcome  $y$  represents verbal reasoning score. Instead of having  $\mathcal{X} \in \mathbb{R}^{2 \times 2 \times 2}$ , we actually have  $\mathcal{X} \in \mathbb{R}^{182 \times 218 \times 182}$  representing each individual's brain MRI scan prior to preprocessing. And the auxiliary covariate vector  $\mathbf{z}$  consists of five variables, including sex, age and indicators of European American, African American and Hispanic American (no subjects are of Asian origin). We note that the element-wise inner product essentially treats the  $182 \times 218 \times 182$  voxels as independent variables. Therefore, model (1) can be also viewed as an ultra-high-dimensional linear model, and it can be extended easily to other types of outcomes through generalized linear models (GLMs).

Similar to matrix decomposition, tensor decomposition is commonly used to obtain a low-rank approximation of the original tensor. There exist different methods of tensor decomposition for different purposes (Kolda and Bader 2009), for example, CPD and Tucker decomposition. For CPD, a rank- $R$  approximation of tensor  $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times \dots \times p_D}$  is represented as

$$\mathcal{B} = \sum_{r=1}^R \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \dots \circ \boldsymbol{\beta}_{r,D}, \quad (2)$$

where  $\circ$  is outer product,  $\boldsymbol{\beta}_{r,d}$  is a  $p_d$ -dimensional vector corresponding to the  $d$ th mode, and  $R$  is defined as the rank if the number of terms  $R$  is minimal. Notice that, Equation (2) is essentially a singular value decomposition when  $D = 2$ , where  $\text{rank}(\mathcal{B}) = R$  and  $\boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2}$  is a rank-1 matrix. When the tensor coefficients  $\mathcal{B}^*$  can be approximated by a CPD as in (2), the imaging regression model (1) becomes

$$y = \mathbf{z}^\top \boldsymbol{\theta}^* + \left\langle \mathcal{X}, \sum_{r=1}^R \boldsymbol{\beta}_{r,1}^* \circ \boldsymbol{\beta}_{r,2}^* \circ \dots \circ \boldsymbol{\beta}_{r,D}^* \right\rangle + \epsilon. \quad (3)$$

Model (3) was first considered by Zhou, Li, and Zhu (2013). Compared to (1), model (3) significantly reduces the dimensionality of tensor components, from  $\prod_d p_d$  in (1) to  $R \sum_d p_d$  in (3). Furthermore, in contrast to model (1) which essentially vectorizes the entire tensor, model (3) is able to keep the inherent spatial structure of the tensor image.

While the CPD based model (3) is appealing, it does not address one important characteristic of imaging coefficients—piecewise smoothness. Piecewise smoothness describes the fact that voxels in close proximity are more likely to have close coefficients. For a two-dimensional image, a commonly used technique to address this issue is through TV (Rudin, Osher, and Fatemi 1992; Rudin and Osher 1994). Given a two-dimensional image (represented by a matrix)  $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2}$ , the TV of  $\mathcal{B}$  takes the form



$$\|\mathcal{B}\|_{\text{TV}} = \sum_{1 \leq j_1 \leq p_1, 1 \leq j_2 \leq p_2} \{|\mathcal{B}_{j_1 j_2} - \mathcal{B}_{j_1 j_2 - 1}| + |\mathcal{B}_{j_1 j_2} - \mathcal{B}_{j_1 - 1 j_2}|\}, \quad (4)$$

where  $\mathcal{B}_{j_1, 0} = \mathcal{B}_{j_1, 1}$  and  $\mathcal{B}_{0, j_2} = \mathcal{B}_{1, j_2}$ . The TV in (4) was referred to the  $\ell_1$ -norm, or anisotropic version of TV. An  $\ell_2$ -norm, or isotropic version of TV could be defined similarly. We note that the TV was originally defined by derivatives for continuous functions from  $\Omega \rightarrow \mathbb{R}$  with  $\Omega \subset \mathbb{R}^{p_1 \times p_2}$ . The definition (4) is the discretized TV. In the work of Wang, Zhu, and Alzheimer's Disease Neuroimaging Initiative (2017), they proposed to minimize the TV of the coefficients  $\mathcal{B}$  in the linear model (1) for two-dimensional image. While their approach works for two-dimensional image, it is nearly impossible to apply their method for the PNC data due to the computational challenges of TV penalized approach for higher order images. This also forms the motivation of this article.

In the remainder of this article, we use calligraphic capital letters to denote tensors, bold capital letters to denote matrices, bold small letters to denote vectors, and unbold letters to denote scalars. Given a tensor  $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times \dots \times p_D}$ , we let  $\mathbf{B}_{(d)} \in \mathbb{R}^{p_d \times \prod_{d' \neq d} p_{d'}}$  be the mode- $d$  matricization of  $\mathcal{B}$  and  $\text{vec}(\mathcal{B})$  be the vectorization of  $\mathcal{B}$ . Specifically, the  $(i_1, \dots, i_D)$  element of tensor  $\mathcal{B}$  maps to the  $(i_d, j)$  element of  $\mathbf{B}_{(d)}$ , with  $j = 1 + \sum_{d' \neq d} (i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$ . The  $(i_1, \dots, i_D)$  element of tensor  $\mathcal{B}$  maps to the  $j'$ th entry of  $\text{vec}(\mathcal{B})$ , with  $j' = 1 + \sum_{d=1}^D (i_d - 1) \prod_{d'=1}^{d-1} p_{d'}$ . We further let  $\|\mathcal{B}\|_F = \sqrt{\text{vec}(\mathcal{B})^\top \text{vec}(\mathcal{B})}$  be the Frobenius norm of  $\mathcal{B}$ . In addition, we let  $\otimes$  represents Kronecker product and recall that  $\circ$  represents the outer product.

## 2.2. The Internal Variation (IV)

To account for the piecewise smooth nature of imaging coefficients, we propose an innovative quantity named IV to substitute TV for higher order tensors. The IV is based on the CPD. Suppose a tensor  $\mathcal{B}$  can be approximated by a low-rank CPD as in (2), the IV of the tensor  $\mathcal{B}$  is defined as

$$\|\mathcal{B}\|_{\text{IV}} = \sum_{r=1}^R \prod_{d=1}^D \|\boldsymbol{\beta}_{r,d}\|_{\text{TV}}, \quad (5)$$

where  $\|\cdot\|_{\text{TV}}$  is the one-dimensional TV:  $\|\mathbf{b}\|_{\text{TV}} = \sum_{j=2}^p |b_j - b_{j-1}|$  for any  $\mathbf{b} \in \mathbb{R}^p$ . The IV (5) sums the products of TVs from all modes over all ranks. The idea of the IV (5) is to approximate the structure of TV, while reducing the dimensionality of tensor components through CPD. When the rank  $R = 1$ , the following connections between IV and TV holds.

**Proposition 1.** Suppose  $\mathcal{B}$  can be decomposed as  $\mathcal{B} = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \dots \circ \boldsymbol{\beta}_D$  for some  $\boldsymbol{\beta}_d \in \mathbb{R}^{p_d}$ ,  $d = 1, \dots, D$ , then:

$$\|\mathcal{B}\|_{\text{IV}} \leq \frac{2^{D-1}}{D} \|\mathcal{B}\|_{\text{TV}}. \quad (6)$$

**Remark 1.** We note that the CPD of a tensor may not be unique due to scaling, permutation and other identifiability issues. However, the IV would stay invariant under most circumstances, thus Proposition 1 would also hold. A detailed discussion on the identifiability issues is deferred to Section 2.5.

More generally, when  $\mathcal{B} = \sum_{r=1}^R \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \dots \circ \boldsymbol{\beta}_{r,D}$  for rank  $r > 1$ , above relationship between IV and TV may not hold, as the outer summation over rank  $r$  breaks the consistency of the inequality. However, it is still beneficial to substitute TV with IV as it provides a smooth solution with minimum computational burden. Given a sample of  $n$  subjects, let  $y_i$ ,  $\mathbf{z}_i$  and  $\mathcal{X}_i$ ,  $i = 1, \dots, n$ , be the sample version of  $y$ ,  $\mathbf{z}$  and  $\mathcal{X}$ , respectively, and let  $\mathbf{y} = (y_1, \dots, y_n)^\top$  and  $\mathbf{Z} = (\mathbf{z}_1^\top, \dots, \mathbf{z}_n^\top)^\top$ . Assume that the linear model (3) holds, then the squared loss can be written as

$$\begin{aligned} L(\boldsymbol{\theta}, \mathcal{B}) &= L(\boldsymbol{\theta}, \boldsymbol{\beta}_{1,1}, \dots, \boldsymbol{\beta}_{R,D}) \\ &\equiv \sum_{i=1}^n \left( y_i - \boldsymbol{\theta}^\top \mathbf{z}_i - \left\langle \mathcal{X}_i, \sum_{r=1}^R \boldsymbol{\beta}_{r,1} \circ \dots \circ \boldsymbol{\beta}_{r,D} \right\rangle \right)^2. \end{aligned} \quad (7)$$

Let  $\nu$  be a smooth tuning parameter, a natural TV constrained minimization problem can be written as

$$\text{minimize } L(\boldsymbol{\theta}, \mathcal{B}), \quad \text{subject to } \|\mathcal{B}\|_{\text{TV}} \leq \nu. \quad (8)$$

When the IV can be upper bounded by TV as in Proposition 1, the solution of the following IV constrained minimization (9) is also a solution of (8):

$$\text{minimize } L(\boldsymbol{\theta}, \mathcal{B}), \quad \text{subject to } \|\mathcal{B}\|_{\text{IV}} \leq \frac{D\nu}{2^{D-1}}. \quad (9)$$

On the other hand, we may also consider the following IV penalized minimization problem

$$\text{minimize } L(\boldsymbol{\theta}, \mathcal{B}) + \lambda \|\mathcal{B}\|_{\text{IV}}. \quad (10)$$

Here,  $\lambda$  is a different smoothing parameter and will be discussed in detail later. We note that the problems (9) and (10) may not be equivalent because the squared loss  $L(\boldsymbol{\theta}, \mathcal{B})$  is not convex on  $\mathcal{B}$ . We shall focus on the IV penalized minimization (10) due to the computational benefits it brings (will show in next subsection). The TV minimization approach for one-dimensional regression (Tibshirani 2014) and two-dimensional imaging regression (Wang, Zhu, and Alzheimer's Disease Neuroimaging Initiative 2017) have been studied intensively. However, the piecewise smooth nature of coefficients has never been considered for high-order images. To the best of our knowledge, the IV penalized estimation is the first computationally feasible approach that producing smooth image coefficients in the form of high-order tensors. We name the IV penalized minimization approach (10) SHIV and we will show the computations of SHIV in the next subsection.

## 2.3. Computation

By adopting a blockwise coordinate descent algorithm, the complicated minimization problem (10) can be solved through a sequence of generalized Lasso problems (Tibshirani and Taylor 2011). To show this, we need a few more notations. For  $r = 1, \dots, R$ , let  $\mathbf{e}_r = (0, \dots, 0, 1, 0, \dots, 0)^\top$  be an  $R$  dimensional vector with the  $r$ th element being one and the rest being zero, let  $\mathbf{A}_r$  be an  $R \times R$  matrix with the  $(r, r)$ th element being one and the rest being zero. For  $d = 1, \dots, D$ , let  $\mathbf{R}_d$  be a  $(p_d - 1) \times p_d$  matrix

$$\mathbf{R}_d = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ & & \vdots & & & \\ 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}_{(p_d-1) \times p_d}.$$

Also, let  $\mathbf{B}_d = [\boldsymbol{\beta}_{1,d}, \dots, \boldsymbol{\beta}_{R,d}]_{p_d \times R}$  be the target parameters at mode  $d$  and  $\mathbf{b}_d = \text{vec}(\mathbf{B}_d) \in \mathbb{R}^{Rp_d}$  be the vectorization of  $\mathbf{B}_d$ . Finally, let  $\mathbf{X}_{i,(d)}$  be the mode- $d$  matricization of  $\mathcal{X}_i$ .

Now we are ready to demonstrate the computation. We alternately update  $\hat{\boldsymbol{\theta}}$  and  $\mathbf{B}_d$ ,  $d = 1, \dots, D$ , while fixing other components. We first note that for the IV,

$$\|\mathcal{B}\|_{\text{IV}} = \sum_{r=1}^R \left( \prod_{d' \neq d} \|\boldsymbol{\beta}_{r,d'}\|_{\text{TV}} \right) \|\boldsymbol{\beta}_{r,d}\|_{\text{TV}} = \|\mathbf{D}_d \mathbf{b}_d\|_1, \quad (11)$$

where

$$\mathbf{D}_d = \sum_{r=1}^R \Lambda_{r,d} \mathbf{A}_r \otimes \mathbf{R}_d \in \mathbb{R}^{(R-1)p_d \times Rp_d}, \quad (12)$$

$$\Lambda_{r,d} = \prod_{\substack{d'=1, \dots, D \\ d' \neq d}} \|\boldsymbol{\beta}_{r,d'}\|_{\text{TV}}.$$

Here, the equality (11) holds for any  $d = 1, \dots, D$ . We note that  $\Lambda_{r,d}$  contains no parameters in  $\boldsymbol{\beta}_{r,d}$ , so the RHS of (11) can be viewed as a weighted  $\ell_1$ -norm of  $\mathbf{b}_d = \text{vec}(\mathbf{B}_d)$ .

For the squared loss, we have

$$\left\langle \mathcal{X}_i, \sum_{r=1}^R \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \dots \circ \boldsymbol{\beta}_{r,D} \right\rangle$$

$$= \sum_{r=1}^R \underbrace{\mathbf{X}_{i,(d)}(\boldsymbol{\beta}_{r,D} \otimes \dots \otimes \boldsymbol{\beta}_{r,d+1} \otimes \boldsymbol{\beta}_{r,d-1} \otimes \dots \otimes \boldsymbol{\beta}_{r,1})}_{\mathbf{m}_{i,r,d}} \boldsymbol{\beta}_{r,d}.$$

Here, we define the  $p_d$ -dimensional vector  $\mathbf{m}_{i,r,d} = \mathbf{X}_{i,(d)}(\boldsymbol{\beta}_{r,D} \otimes \dots \otimes \boldsymbol{\beta}_{r,d+1} \otimes \boldsymbol{\beta}_{r,d-1} \otimes \dots \otimes \boldsymbol{\beta}_{r,1})$ . We note that  $\mathbf{m}_{i,r,d}$  does not contain parameters in  $\boldsymbol{\beta}_{r,d}$ , so the RHS of the above equality can be viewed as a linear combination of  $\boldsymbol{\beta}_{r,d}$ . If we further allow

$$\mathbf{M}_d = (\mathbf{M}_{d,1}^\top, \dots, \mathbf{M}_{d,n}^\top)^\top \in \mathbb{R}^{n \times (p_d R)}, \quad (13)$$

$$\mathbf{M}_{d,i} = \left( \sum_{r=1}^R \mathbf{e}_r \otimes \mathbf{m}_{i,r,d} \right)^\top,$$

and  $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\boldsymbol{\theta}$ , then the squared loss in (7) can be written as

$$L(\boldsymbol{\theta}, \mathcal{B}) = \frac{1}{2n} \|\tilde{\mathbf{y}} - \mathbf{M}_d \mathbf{b}_d\|_2^2. \quad (14)$$

Combining (14) and (11), the criterion function (10) can be transformed to the following penalized estimation given  $\mathbf{B}_1, \dots, \mathbf{B}_{d-1}, \mathbf{B}_{d+1}, \dots, \mathbf{B}_D$  and  $\boldsymbol{\theta}$ :

$$\hat{\mathbf{b}}_d \in \underset{\mathbf{b}_d \in \mathbb{R}^{Rp_d}}{\text{argmin}} \frac{1}{2n} \|\tilde{\mathbf{y}} - \mathbf{M}_d \mathbf{b}_d\|_2^2 + \lambda \|\mathbf{D}_d \mathbf{b}_d\|_1. \quad (15)$$

The minimization (15) is essentially a generalized Lasso and can be further transformed to a standard Lasso problem (Tibshirani and Taylor 2011). Now the computation is standard, for example, the coordinate descent algorithm can be applied.

Given  $\hat{\mathcal{B}}$ , the estimation of  $\boldsymbol{\theta}$  is straightforward. Let  $\mathbf{y}' = (y'_1, \dots, y'_n)$  with  $y'_i = y_i - \langle \mathcal{X}_i, \hat{\mathcal{B}} \rangle$ . We obtain  $\boldsymbol{\theta}$  by least square estimation

$$\hat{\boldsymbol{\theta}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}'. \quad (16)$$

As a summary, we formulate the detailed computation algorithm as in Algorithm 1. To simplify the notation in the algorithm, let  $\boldsymbol{\alpha} = \text{vec}(\boldsymbol{\theta}, \mathbf{B}_1, \dots, \mathbf{B}_D)$ . Algorithm 1 is a blockwise coordinate descent algorithm, where  $(\boldsymbol{\theta}, \mathbf{B}_1, \dots, \mathbf{B}_D)$  is estimated cyclically. In practice, the initial value for  $\boldsymbol{\theta}$  is typically assigned as the ordinary least square result of  $\mathbf{y}$  against  $\mathbf{z}$ , while the initial values for  $(\mathbf{B}_1, \dots, \mathbf{B}_D)$  can be chosen from independent random Gaussian matrices with mean 0 and small variance, or estimation results of Zhou, Li, and Zhu (2013) without penalization.

---

**Algorithm 1** A blockwise coordinate descent algorithm for SHIV

---

**Input:** response  $y_i$ , tensor image  $\mathcal{X}_i$ , and auxiliary covariate  $\mathbf{z}_i$  for subject  $i = 1, \dots, n$ , tensor rank  $R$ , penalization coefficient  $\lambda$ , number of iterations  $t = 1$ , initial value  $\boldsymbol{\alpha}^{(0)}$ , convergence threshold  $\tau$ , and maximum number of iterations  $s_{\max}$ .

**repeat**

**for**  $d = 1, \dots, D$  **do**

    Estimate  $\mathbf{B}_d^{(t)}$  via minimizing (15) with  $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\boldsymbol{\theta}^{(t)}$ ,  $\mathbf{M}_d$  given in (13) and  $\mathbf{D}_d$  given in (12).

**end for**

  Estimate  $\boldsymbol{\theta}^{(t)}$  through (16)

$t = t + 1$

**until**  $s > s_{\max}$  **or**  $\frac{L(\boldsymbol{\alpha}^{(t-1)}) - L(\boldsymbol{\alpha}^{(t)})}{L(\boldsymbol{\alpha}^{(t-1)})} < \tau$ .

---

## 2.4. Model Selection and Degrees of Freedom

In the literature, the tuning parameter  $\lambda$  is usually chosen by certain model selection criteria, such as Bayesian information criterion (BIC), Akaike information criterion (AIC) or Mallows's  $C_p$  criterion. All of these criteria employ degrees of freedom to assess the risk. The degrees of freedom essentially describe the effective number of parameters used by a fitting procedure. Although the concept of degrees of freedom is well understood in linear models, it is not the case in our problem. We note that model (3) is a degree  $D$  polynomial of parameters, the degrees of freedom of the fit in the SHIV approach need to be carefully derived. This is done in this subsection.

Suppose that the tensor images are observed without error. For a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , the number of degrees of freedom of  $f$  is defined as

$$\text{df}(f) = \frac{1}{\sigma^2} \sum_{i=1}^n \text{cov}(f_i(\mathbf{y}), y_i),$$

where  $f_i$  is the  $i$ th coordinate function of  $f$ . In our problem, given  $\lambda$ ,  $f(\cdot; \lambda)$  is given by

$$f_i(\mathbf{y}; \lambda) = \hat{\boldsymbol{\theta}}^\top \mathbf{z}_i + \left\langle \sum_{r=1}^R \hat{\boldsymbol{\beta}}_{r,1}(\mathbf{y}; \lambda) \circ \hat{\boldsymbol{\beta}}_{r,2}(\mathbf{y}; \lambda) \circ \dots \circ \hat{\boldsymbol{\beta}}_{r,D}(\mathbf{y}; \lambda), \mathcal{X}_i \right\rangle. \quad (17)$$

The following proposition provides the degrees of freedom of  $f(\cdot; \lambda)$ .

**Proposition 2.** Under usual regularity conditions (see appendix for details) the fit (17) with fixed penalty level  $\lambda$  has the following degrees of freedom

$$\text{df}(f) = q + \sum_{d=1}^D \sum_{r=1}^R \mathbb{E}[\text{number of fused groups in } \hat{\beta}_{r,d}(\mathbf{y}; \lambda)].$$

By Proposition 2, we see that the degrees of freedom for the fit of IV penalization is equal to the sum of number of fused groups in  $\hat{\beta}_{r,d}$  across all ranks and modes. We note that Proposition 2 may be also generalized to other forms of penalties in tensor regression, for example, Lasso in Zhou, Li, and Zhu (2013). This finding is important as BIC, AIC, or Mallows's  $C_p$  criteria can all be directly applied for model selection. In our simulation and real brain image regression example, the penalty level  $\lambda$  is chosen by the following BIC criterion:

$$\text{BIC}(\lambda) = \sum_{i=1}^n \left( y_i - f_i(\mathbf{y}; \lambda) \right)^2 + \text{df}(f) \cdot \log n.$$

## 2.5. Identifiability Issues

Identifiability issues arise commonly in tensor decomposition. On the one hand, nonidentifiability in CPD is attributed to two elementary indeterminacies, namely scaling and permutation. On the other hand, the CPD of a tensor might be represented by more than one combination of rank-1 tensors even after the aforementioned indeterminacies when  $D \geq 2$ . When  $D = 2$ , it is clear that a standard singular value decomposition can be imposed. Therefore, we mainly discuss the case when  $D \geq 3$ .

Recall in (2) that a tensor  $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times \dots \times p_D}$  under CPD can be represented by  $\mathbf{B}_1, \dots, \mathbf{B}_D$ . For convenience, here we use  $\mathcal{B} = \llbracket \mathbf{B}_1, \dots, \mathbf{B}_D \rrbracket$  to indicate the CPD. If  $\text{rank}(\mathcal{B}) = R$ , then a scaling indeterminacy states that

$$\mathcal{B} = \llbracket \mathbf{B}_1, \dots, \mathbf{B}_D \rrbracket = \llbracket \mathbf{B}_1 \Gamma_1, \dots, \mathbf{B}_D \Gamma_D \rrbracket,$$

where  $\Gamma_d = \text{diag}(\gamma_{1,d}, \dots, \gamma_{R,d})$ ,  $d = 1, \dots, D$ , are diagonal matrices satisfying  $\prod_{d=1}^D \gamma_{r,d} = 1$  for each  $r$ . Meanwhile, a permutation indeterminacy states that

$$\mathcal{B} = \llbracket \mathbf{B}_1, \dots, \mathbf{B}_D \rrbracket = \llbracket \mathbf{B}_1 \Omega, \dots, \mathbf{B}_D \Omega \rrbracket,$$

where  $\Omega$  is an arbitrary permutation matrix. Nevertheless, we note that the IV is invariant under scaling and permutation indeterminacies. That is to say,

$$\begin{aligned} \|\llbracket \mathbf{B}_1, \dots, \mathbf{B}_D \rrbracket\|_{\text{IV}} &= \|\llbracket \mathbf{B}_1 \Gamma, \dots, \mathbf{B}_D \Gamma \rrbracket\|_{\text{IV}} \\ &= \|\llbracket \mathbf{B}_1 \Omega, \dots, \mathbf{B}_D \Omega \rrbracket\|_{\text{IV}}. \end{aligned}$$

To avoid complications, we shall adopt a specific parameterization to handle these two indeterminacies. Let  $\lambda_r = \prod_d \|\beta_{r,d}\|_2$  denote the  $r$ th “eigenvalue” that corresponds to the weight at the  $r$ th rank-1 tensor. To address the scaling indeterminacy, we let  $\gamma_{r,d} = (\lambda_r)^{1/d} / \|\beta_{r,d}\|_2$  at each  $r$ , such that each mode shares the same weight. Meanwhile, we permute the columns of  $\mathbf{B}_1, \dots, \mathbf{B}_D$  simultaneously, such that  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_R$ , which addresses the permutation indeterminacy in probability.

Next, we provide sufficiency and necessity conditions to address nonidentifiability beyond scaling and permutation issues. We first introduce the concept of  $k$ -rank (Kruskal 1977), denoted by  $K_A$  for an arbitrary matrix  $\mathbf{A}$ , that is,

$$K_A = \max\{k : \text{any } k \text{ columns of } \mathbf{A} \text{ are linearly independent}\}.$$

Theorem 3 of Sidiropoulos and Bro (2000) and Corollary 1 of Liu and Sidiropoulos (2001) provided the sufficiency and necessity conditions for the identifiability of a general tensor, respectively, which is also applicable to the tensor  $\mathcal{B}$  in this article:

**Proposition 3.** Suppose  $\mathcal{B}$  has rank  $R$  and  $\mathcal{B} = \sum_{r=1}^R \beta_{r,1} \circ \beta_{r,2} \circ \dots \circ \beta_{r,D}$ . Then the following two conditions provide respective sufficiency and necessity for the uniqueness of tensor  $\mathcal{B}$  up to scaling and permutation:

1. Sufficiency (Sidiropoulos and Bro 2000):  $\sum_{d=1}^D K_{B_d} \geq 2R + (D - 1)$ ,
2. Necessity (Liu and Sidiropoulos 2001):  $\min_{d=1, \dots, D} \left\{ \prod_{d' \neq d} \text{rank}(\mathbf{B}_{d'}) \right\} \geq R$ .

Ten Berge and Sidiropoulos (2002) also demonstrated that the sufficiency condition becomes necessary when  $R = 2$  or 3. Proof of Proposition 3 is omitted as it is a direct consequence of Sidiropoulos and Bro (2000) and Liu and Sidiropoulos (2001). In general, the sufficiency condition is a practical tool to check the uniqueness of the desired tensor  $\mathcal{B}$ .

## 3. Simulation Studies

We conduct a comprehensive simulation study to evaluate the estimation performance of our approach. The proposed method is compared with three competing methods including the tensor regression (TR) and the tensor regression with Lasso (TR Lasso) proposed by Zhou, Li, and Zhu (2013), and the voxelwise regression (VoxelReg) commonly used by practitioners, where linear regression is fitted using the response variable against each voxel.

We investigate the influence of the sample size and the shape of  $\mathcal{B}$  on the recovery of the values of  $\mathcal{B}$ . The coefficient tensor  $\mathcal{B}$  in the studies varies but is assumed to be sparse and has a support representing a certain shape. This simulates the effect that only a subregion of brain has influence on a phenotype. Specifically, we assume the sample size  $n = 200, 350, 500$ , and 700. The shape of  $\mathcal{B}$  is specified as “one brick,” “two bricks,” “three-dimensional-cross,” or “pyramid.” We let the size of  $\mathcal{B}$  be  $30 \times 30 \times 30$ . For “one brick,” we assume the signal to be a cube with the length of each edge equal to 5, and the frontal-bottom-left corner is located at (15, 15, 20). That is,  $\mathcal{B}_{15:19,15:19,20:24} = 1$ , and the rest of the elements of  $\mathcal{B}$  equal to 0. And it is easy to check that  $\text{rank}(\mathcal{B}) = 1$ . Similarly, for “two bricks,” we assume  $\mathcal{B}_{15:19,15:19,20:24} = 1$  and  $\mathcal{B}_{20:24,20:24,25:29} = 1$ , and  $\text{rank}(\mathcal{B}) = 2$ . For “three-dimensional-cross,” we assume  $\mathcal{B}_{15:26,19:22,19:22} = 1$ ,  $\mathcal{B}_{19:22,15:26,19:22} = 1$ , and  $\mathcal{B}_{19:22,19:22,15:26} = 1$ , which lead to  $\text{rank}(\mathcal{B}) = 3$ . For “pyramid,” we assume that each layer is a square with the top layer being a single pixel, and that each layer has two more pixels along each side than the layer above. We let

**Table 1.** (Simulation results) Mean and standard error of  $\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F$  using SHIV and competing methods under different signal shapes of  $\mathcal{B}$  and sample size  $n$ .

Signal shape	$n$	VoxelReg	TR	TR Lasso	SHIV
One brick	200	132.75 (6.77)	15.33 (6.28)	12.34 (6.10)	<b>4.24</b> (6.20)
	350	99.69 (3.85)	2.37 (4.73)	2.14 (4.23)	<b>0.50</b> (1.71)
	500	82.84 (2.54)	0.53 (0.93)	0.58 (0.84)	<b>0.20</b> (0.03)
	700	69.93 (1.92)	0.38 (0.03)	0.51 (0.06)	<b>0.16</b> (0.03)
Two bricks	200	186.93 (10.09)	34.55 (2.85)	21.38 (1.56)	<b>6.38</b> (5.34)
	350	140.57 (5.15)	26.63 (1.63)	21.38 (3.32)	<b>4.56</b> (7.06)
	500	117.34 (3.74)	9.93 (8.84)	6.86 (6.89)	<b>0.87</b> (2.65)
	700	99.09 (2.78)	1.00 (2.33)	1.07 (1.94)	<b>0.25</b> (0.03)
Three-dimensional cross	200	250.61 (12.95)	42.42 (3.05)	30.86 (1.35)	<b>28.73</b> (6.15)
	350	188.11 (6.66)	43.93 (3.31)	32.45 (4.40)	<b>9.35</b> (7.90)
	500	156.59 (4.89)	24.96 (10.71)	15.44 (11.91)	<b>3.86</b> (5.61)
	700	132.33 (3.40)	0.99 (1.14)	1.47 (0.15)	<b>0.55</b> (0.17)
Pyramid	200	305.56 (15.13)	87.09 (13.69)	34.56 (1.24)	<b>32.65</b> (6.43)
	350	231.59 (8.25)	54.96 (3.32)	33.62 (4.38)	<b>26.21</b> (5.35)
	500	193.33 (6.42)	51.74 (2.50)	29.96 (8.78)	<b>17.06</b> (2.93)
	700	162.73 (4.37)	54.73 (4.19)	18.84 (11.41)	<b>10.66</b> (2.10)

NOTE: Here,  $\mathcal{X}$  is drawn from iid normal distributions. The number marked in bold indicates the best performance in each setting.

the top layer be  $\mathcal{B}_{22,22,27} = 1$  and assume a total of 8 layers<sup>2</sup>. Then  $\mathcal{B}_{22-k:22+k,22-k:22+k,27-k} = 1$  with  $k = 0, \dots, 7$ , and  $\text{rank}(\mathcal{B}) = 8$ . The image  $\mathcal{X}$  was drawn from independent and identically distributed normal distributions. A more realistic case where  $\mathcal{X}$  is real three-dimensional brain images will be considered in the next section.

For the auxiliary covariate, we assume that each row of  $\mathbf{Z}$  satisfies  $\mathbf{Z}_i \stackrel{\text{iid}}{\sim} N(0, I_5)$  and  $\boldsymbol{\theta} = (1, 1, 1, 1, 1)'$ . The response  $y_i$  is simulated as in model (1), with  $\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$  and  $\sigma = 1$ . For each method, we assume the true rank of  $\mathcal{B}$  is known. The tuning parameter  $\lambda$  is selected based on a grid search ranging from 0.05 to 0.09. The BIC is utilized as the selection criterion with degrees of freedom determined by Proposition 2 in Section 2.4. The selection of  $\lambda$  for TR Lasso is based on the default setting of Zhou's Matlab package. We let the TR estimated  $\mathbf{B}_1, \dots, \mathbf{B}_D$  as the initial values of TR Lasso and SHIV. The simulation result is based on 200 replications.

The performance of each method is measured by  $\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F$  where  $\hat{\mathcal{B}}$  is the estimated coefficient tensor, and  $\|\cdot\|_F$  represents the Frobenius norm. The summary of results are presented in Table 1. It can be seen that the proposed SHIV method uniformly outperforms the competing methods through incorporating IV. For each signal shape, the performance of each

method improves as the sample size increases. In addition, we shall notice that the improvement of SHIV due to sample size increase is particularly significant. Furthermore, each method performs worse with the increase of the true rank from “one brick” to “pyramid.” For VoxelReg, since the total number of voxels is much greater than the sample size, each linear regression fitting against one voxel has estimated coefficient far from the true value. This leads to a large error. The results of TR and TR Lasso are similar when the true rank is small, while the penalty term helps improve the estimation accuracy when the true rank increases. In the supplementary materials, the prediction errors in terms of the estimated response for this simulation and the simulation in Section 4.1 are also provided, where SHIV also demonstrate the best performance in terms of out-of-sample testing errors in most settings. Meanwhile, we also provide additional simulation studies where the error term is non-Gaussian. SHIV also shows the best performance.

#### 4. The PNC Study

The PNC (Gur et al. 2012; Satterthwaite et al. 2014) is a large research initiative which collects cognitive, neuroimaging, and genetic data from over 9500 individuals from the greater Philadelphia area with ages ranging from 8 to 21 years.

Among the PNC participants, 1445 children and adolescents participated in the T1-weighted imaging scan. Preprocessing of the raw MRI images are conducted prior to statistical analysis, which include reorientation, bias-field correction, brain extraction, intensity normalization, and linear registration. Here reorientation represents the process of reorienting the MRI images to the standard MNI orientation. Bias-field correction corrects for RF/B1 inhomogeneities (i.e., spatial intensity variations) through a hidden Markov random field model (Zhang, Brady, and Smith 2001). The brain extraction is based upon transforming a standard-space binarized brain mask to the corrected input image (Smith 2002). And intensity normalization brings the intensities across images from all individuals to a common scale. For linear registration, a global optimization method is applied, which uses trilinear interpolation with the correlation ratio as a cost function (Jenkinson and Smith 2001; Jenkinson et al. 2002). Specifically, it combines Powell's conjugate direction method (Powell 1964) with an initial search for the correct orientation. The MNI152 standard-space T1-weighted average structural image is adopted as the template. All the preprocessing steps are implemented via the brain image analysis software FSL (Smith et al. 2004; Jenkinson et al. 2012). MRI images from 997 participants are successfully preprocessed and registered.

One of the limitations in analyzing real imaging data is that we do not know what truly happened to the brain. In Section 4.1, therefore, we use real three-dimensional brain images from the PNC as  $\mathcal{X}$  but use a simulated  $\mathcal{B}$  (and  $\mathbf{y}$ ) with the signal regions in  $\mathcal{B}$  being known. We then compare SHIV with the competing methods and see if each method is able to locate the true signal region. Similar to Section 3, the coefficient tensor  $\mathcal{B}$  is assumed to have a support representing a certain shape, which simulates the effect that only a subregion of brain has influence on a phenotype.

<sup>2</sup>For  $n = 200$ , the top two layers are removed to avoid large parameter size.



**Table 2.** (Real MRI images with hypothetical signals) Mean and standard error of  $\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F$  using SHIV and competing methods under different signal shapes of  $\mathcal{B}^*$  and image resolutions.

Signal shape	Image resolution	VoxelReg	TR	TR Lasso	SHIV
One brick	$47 \times 58 \times 47$	$1.23 \times 10^4$ (38.45)	1.63 (0.20)	<b>0.31</b> (0.07)	0.57 (0.16)
	$57 \times 70 \times 57$	$1.11 \times 10^4$ (59.82)	2.73 (8.13)	<b>0.38</b> (0.43)	1.68 (0.22)
	$71 \times 87 \times 71$	$2.48 \times 10^4$ (77.12)	6.96 (21.28)	<b>0.58</b> (0.73)	1.94 (0.22)
Two bricks	$47 \times 58 \times 47$	$2.60 \times 10^4$ (78.07)	89.85 (34.80)	25.58 (20.19)	<b>12.80</b> (14.88)
	$57 \times 70 \times 57$	$2.43 \times 10^4$ (113.06)	190.22 (47.98)	64.46 (43.72)	<b>22.77</b> (23.74)
	$71 \times 87 \times 71$	$4.96 \times 10^4$ (159.02)	515.19 (98.89)	309.65 (102.97)	<b>65.94</b> (47.36)
Three-dimensional cross	$47 \times 58 \times 47$	$2.08 \times 10^5$ (538.19)	347.57 (83.62)	180.22 (76.01)	<b>32.84</b> (20.01)
	$57 \times 70 \times 57$	$2.12 \times 10^5$ (651.10)	921.12 (264.80)	742.51 (286.51)	<b>48.33</b> (32.67)
	$71 \times 87 \times 71$	$2.32 \times 10^5$ (855.56)	2411.8 (517.24)	2165.0 (544.76)	<b>61.25</b> (40.28)
Pyramid	$47 \times 58 \times 47$	$6.312 \times 10^4$ (215.72)	1033.7 (174.48)	67.23 (56.08)	<b>26.62</b> (11.13)
	$57 \times 70 \times 57$	$9.46 \times 10^4$ (308.70)	2559.2 (854.33)	83.64 (39.44)	<b>68.90</b> (35.56)
	$71 \times 87 \times 71$	$9.56 \times 10^4$ (371.12)	1581.7 (221.72)	438.26 (120.17)	<b>74.99</b> (68.90)

NOTE: Here,  $\mathcal{X}$  represents the real MRI images from the PNC study. The number marked in bold indicates the best performance in each setting.

In Section 4.2, real data analysis is conducted. Both the three-dimensional brain images and the response variable are from the PNC study. For the response variable, we investigate each individual's verbal reasoning ability, which is known as a reliable measure of individuals' general intellectual abilities (Gur et al. 2010), through the Pennsylvania Verbal Reasoning Test. Here the Pennsylvania Verbal Reasoning Test consists of 15 verbal analogy problems with simplified instructions and vocabulary (Gur et al. 2001), where the verbal analogy problems are from the factor-referenced cognitive test kit provided by the Educational Test Service. The total test score is the number of correctly answered problems, which ranges from 0 to 15 and quantifies each individual's verbal reasoning ability. In addition to brain imaging, a total of five covariates are also considered in Section 4.2, including sex, age and race/ethnicity. (Three indicators are adopted for race/ethnicity, which represent European American, African American and Hispanic American. No subjects are of Asian origin.)

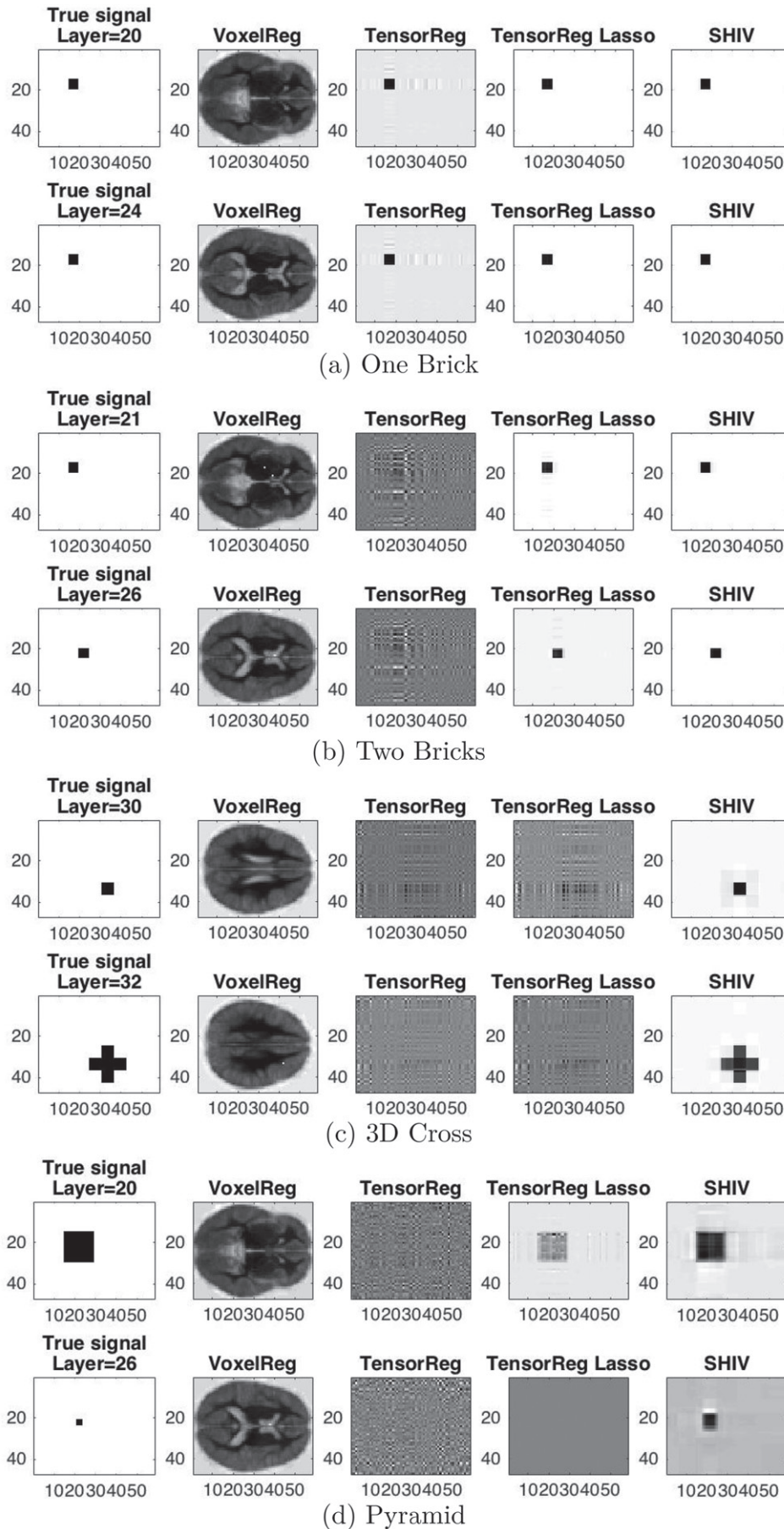
#### 4.1. Real MRI Images With Hypothetical Signals

The size of the MRI images after preprocessing is  $182 \times 218 \times 182$  for each participant. To improve the computational speed, we crop each image such that at each tensor mode, layers of voxels with zero intensity are removed, reducing the size of the images to  $142 \times 175 \times 142$ . The auxiliary covariate  $\mathbf{Z}$ , its coefficient  $\boldsymbol{\theta}$ , the shape of the coefficient  $\mathcal{B}$ , and the simulated noise  $\varepsilon$  are generated in the same way as in Section 3. To examine the performance of each method under different image resolutions, we conduct interpolation to change the size of each image to roughly 33%, 40%, and 50% of its original size; that is, the final

size of each image is reduced to  $47 \times 58 \times 47$ ,  $57 \times 70 \times 57$ , and  $71 \times 87 \times 71$ , respectively. To guarantee the quality of the images, we consider only 497 subjects who have a valid and greater-than-100 standardized Wide Range Achievement Test (WRAT) score. Another reason for such a screening is to improve the computational speed for this study. Furthermore, each image voxel is centralized and standardized across the  $n = 497$  selected subjects.

The shape and strength of the signals are the same as in Section 3, that is, "one brick," "two bricks," "three-dimensional-cross," or "pyramid." The auxiliary covariates are also the same as in Section 3. The simulation results are summarized in Table 2, which was based on 200 independent replicates. It can be seen that every method performs better when the image resolution is small and when the signal shape is simple (e.g., "one brick" and "two bricks"). In particular, TR Lasso has the best performance when the signal shape is "one brick," and SHIV has the best performance for the more complex signal shapes regardless of the image resolution. Meanwhile, the error of VoxelReg is much larger than the remainder of the competing methods. The values are also much larger than their counterparts in Section 3. This is because many voxels are highly correlated, and the regression of  $y$  against each voxel reflects not only the coefficient of that voxel, but multiple correlated voxels simultaneously.

Figure 1 displays how well each method detects signals. Each figure demonstrates two representative layers. It can be seen that in Figure 1(a), TR, TR Lasso and SHIV can clearly recognize the true "one brick" signal. For "two bricks" as illustrated in Figure 1(b), only can TR Lasso and SHIV still recognize the true signal. For "three-dimensional-cross" as illustrated in Figure 1(c), only can SHIV recognize the shape of the true signal. And for the "pyramid" as in Figure 1(d), TR Lasso is able to

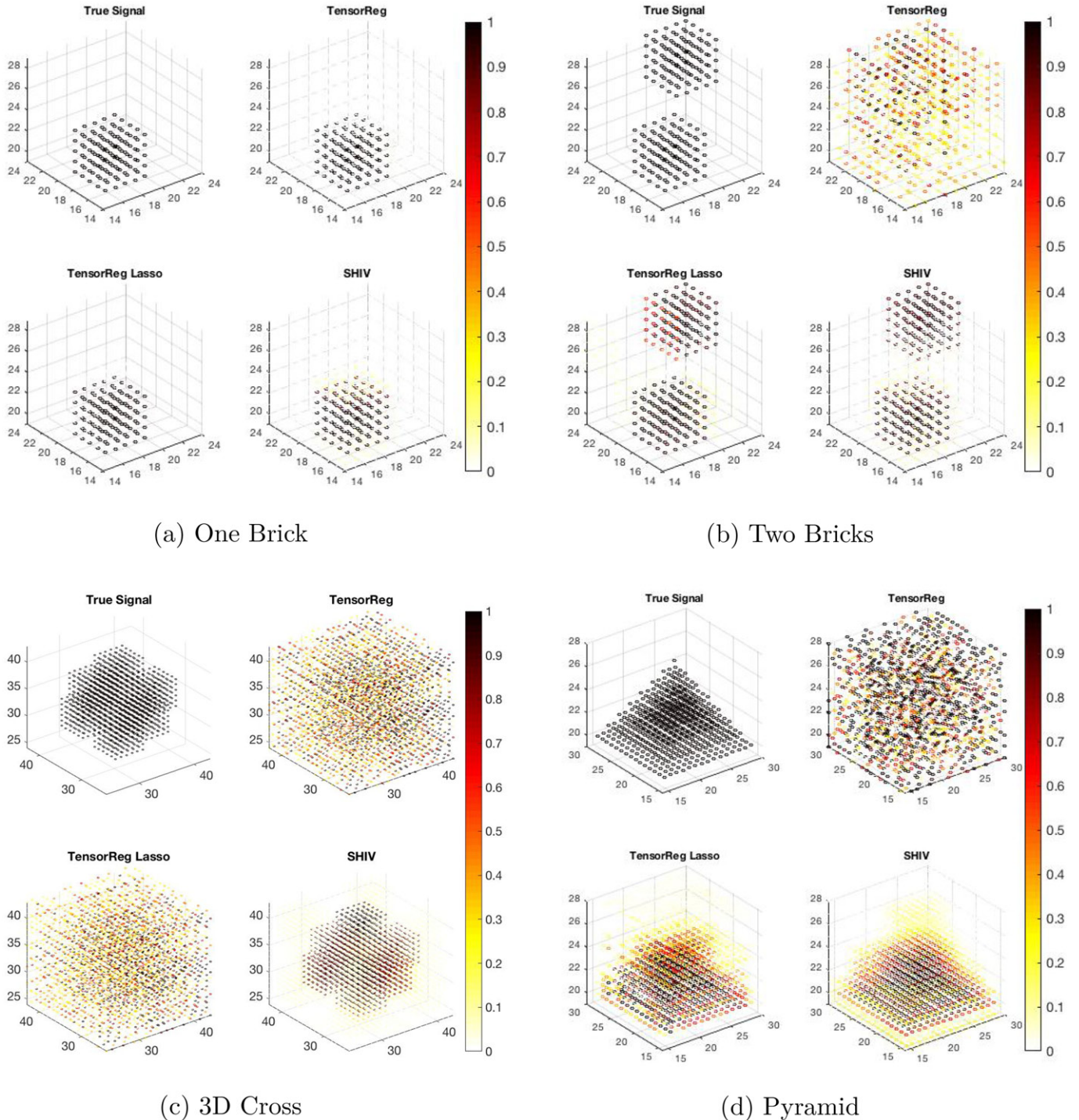


**Figure 1.** (Real MRI images with hypothetical signals) Two horizontal sections of the real three-dimensional image coefficients and estimated coefficients by SHIV and the competing methods are illustrated. VoxelReg only reflects the rough scan of brain layers, and is not able to capture the true signal.

recognize the base part, while the proposed method is still able to roughly capture the top. In contrast, voxelwise regression reflects approximately the brain scan of each layer, and is not able to capture the true signal.

To further compare the proposed method with TR and TR Lasso, we use Figure 2 to reflect the three-dimensional reconstruction of the true signal. Since voxelwise regression is not able to reconstruct the true signal, its three-dimensional reconstruction is not provided. It can be seen that for “one brick” in

Figure 2(a), TR, TR Lasso and SHIV all have good reconstruction. Here, TR and TR Lasso have estimated nonzero values of  $\beta$  closer to the true signal value, reflected by the uniformly black dots. In Figures 2(b) and 2(d), TR is no longer able to reconstruct the true signal shape, whereas the signal shape is visible from the TR Lasso and SHIV methods. In particular, SHIV has better reconstruction than TR Lasso for the “pyramid” shape as illustrated in Figure 2(d). And in Figure 2(c), SHIV is the only method that can reconstruct the true signal shape.



**Figure 2.** (Real MRI images with hypothetical signals) Real three-dimensional image coefficients and the estimated coefficients by SHIV and other competing methods are illustrated.



## 4.2. Neurodevelopmental Data Analysis

The goal of the real data study in this subsection is to detect brain regions that may have association with individuals' verbal reasoning ability. Here,  $n = 978$  individuals who participated in both the MRI scan and the Pennsylvania Verbal Reasoning Test. The images are resized to  $71 \times 87 \times 71$  and standardized as in Section 4.1. To sufficiently explain the data variation, we set the rank  $R$  for each method to range from 7 to 12. For the proposed method and TR Lasso, the tuning parameter  $\lambda = 0.75 \times 10^{-3}$  is selected from  $\{0.50 \times 10^{-3}, 0.75 \times 10^{-3}, \dots, 2.5 \times 10^{-3}\}$ . BIC is utilized to select the tuning parameter as in previous studies.

Figure 3 illustrates the sagittal, coronal and horizontal sections of the subregion identified by SHIV, respectively. While the odd rows demonstrate the absolute magnitude of the coefficient at each voxel, the even rows demonstrate coefficients with large magnitudes (seven standard deviations greater than the mean).

Figure 3 indicates that SHIV can identify clear and strong signals in subregions of the brain as marked by red voxels in the last column. In contrast, the other three methods locate voxels that appear sporadic. Particularly, the voxels identified by TR and TR Lasso spread so widely and their signals appear so weak that we can hardly recognize their existence. If we look closely, we can see that the regions identified by SHIV contain some isolated voxels that are also identified by the other three methods. Therefore, there is a certainty of consistency among themselves while SHIV is the only method that presents clear and convincing patterns.

It is noteworthy that the subregion with the strongest signal identified by SHIV is located across the cingulate cortex and the corpus callosum. The anterior cingulate cortex has been identified as being involved in vocalization of emotions and socially driven interactions (Etkin, Egner, and Kalisch 2011; Lavin et al. 2013), the former of which resulted from its connection with vocalization areas in the frontal lobes, namely the Broca's area. It is also reported that both the anterior and the posterior cingulate cortex are associated with cognition (Sanfey et al. 2003; Leech and Sharp 2013). Meanwhile, subregions of corpus callosum are reported as being associated with dyslexia (Hynd et al. 1995; Von Plessen et al. 2002). Nevertheless, the identified subregion provided by SHIV, which is across two human brain anatomical regions, has not been reported by the existing literature to the best of our knowledge.

In the top right panel of Figure 3, the identified red subregion has a total of 48 voxels. Three matrices, namely  $\hat{B}_{39,22:30,41:49}$ ,  $\hat{B}_{40,22:30,41:49}$ , and  $\hat{B}_{41,22:30,41:49}$ , corresponding to this subregion of the coefficient tensor identified by SHIV is provided in the supplementary materials. The average of the coefficients in this subregion is 0.0232. This indicates that if the imaging intensity of the entire identified subregion is increased by 1, individuals' average verbal reasoning test score will increase by about  $0.0232 \times 48 \approx 1.114$  (out of 15) points. In fact, a linear regression that fits the verbal reasoning ability against the average voxel intensity within the identified subregion revealed a  $p$ -value of  $3.48 \times 10^{-4}$  for the regression coefficient, after controlling for the other covariates, supporting the significant association between the identified subregion and the verbal reasoning ability.

Furthermore, we conduct robustness check to examine the sensitivity of SHIV against the change of the smoothing parameter  $\lambda$ . In Figure 4, we illustrate the estimated coefficient tensor  $\hat{B}$  at the identified subregion (which is in contrast to the absolute values illustrated in Figure 3). Here blue represents negative coefficients, while yellow and red represent positive coefficients. It can be seen that the location and the sign of the identified subregions are almost invariant against the change of  $\lambda$ , indicating the robustness of the proposed method. Meanwhile, the volume of the identified subregion grows with the increase of  $\lambda$ . And with a larger  $\lambda$ , the estimated coefficient tensor  $\hat{B}$  tends to have bigger "blocks" which share the same or similar values. In addition, we conduct sensitivity analysis with respect to data reduction. Specifically, 5%, 10%, and 15% of the PNC data are randomly removed, while other parameters are held the same. The results, provided as Figure S1 in the supplementary materials, show that the identified regions are also robust against data reduction.

In Figure 4, it is also important to note that the identified red subregion in the cingulate cortex and corpus callosum is surrounded by blue color, regardless of the  $\lambda$ . That is, the reduction of the boundaries of the identified subregion, which to some extent indicates the shrinkage of the volume of such subregion, may lead to deterioration of individuals' verbal reasoning ability.

## 5. Discussion

This work focuses on how brain-imaging data may be used to understand intellectual disabilities. We propose SHIV, an innovative framework to impose the piecewise-smooth structure of imaging coefficients in high-order imaging regression problem. This sets the stage to further examine the genetic effects and genetic-imaging interactions on the intellectual disabilities or other complex diseases. This task is extremely difficult due to the high dimensionalities of both genetic and imaging data, and general lack of knowledge on how these data may contribute to the intellectual disabilities. For example, Mehta, Gruen, and Zhang (2017) integrated a region-of-interest (ROI) based score into genetic models to analyze learning disorders. However, while it is a sensible preliminary step, the ROIs based approach may not take full advantage of the rich information in the imaging data. Thus, an approach that is able to integrate the high-dimensional genetic effects into the voxel-wise imaging model would be very useful.

As the verbal reasoning score is continuous in the PNC data, linear scalar-on-image model was applied in this article. However, the IV and the IV penalized estimation can be easily generalized to GLM. In a GLM, Suppose

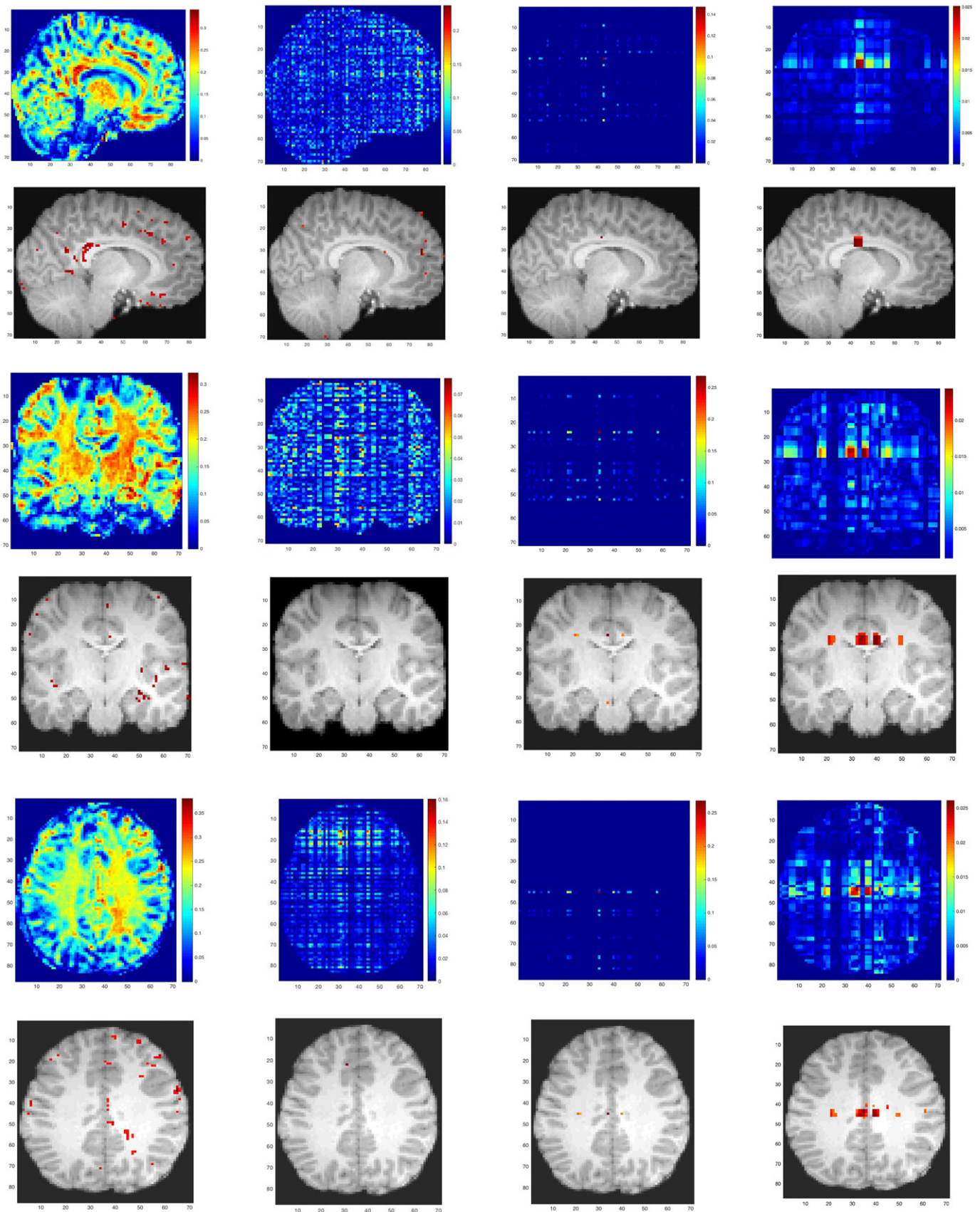
$$g(\mathbb{E}(\mathbf{y})) = \mathbf{z}^\top \boldsymbol{\theta} + \left\langle \mathcal{X}, \sum_{r=1}^R \boldsymbol{\beta}_{r,1}^* \circ \boldsymbol{\beta}_{r,2}^* \circ \dots \circ \boldsymbol{\beta}_{r,D}^* \right\rangle$$

with  $g(\cdot)$  being certain link functions. Then the SHIV estimates can be modified to minimize the IV penalized negative likelihood:

$$L(\mathcal{B}) = -(1/n)\mathbb{P}(\mathbf{y}|\mathcal{X}, \mathbf{Z}, \mathcal{B}) + \lambda \|\mathcal{B}\|_{\text{IV}}.$$

The computation of IV penalized GLM could also be carried out similarly as in Section 2.3.





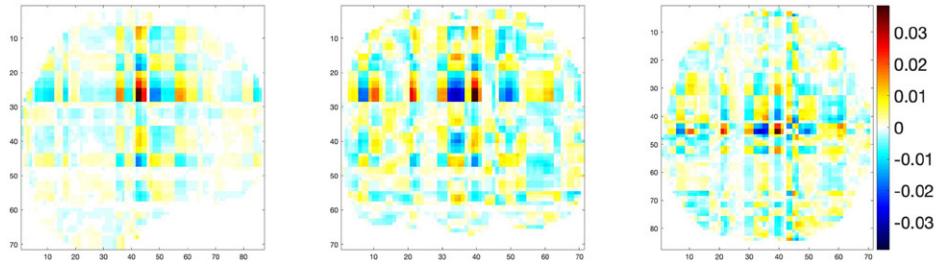
(a) VoxelReg

(b) TensorReg

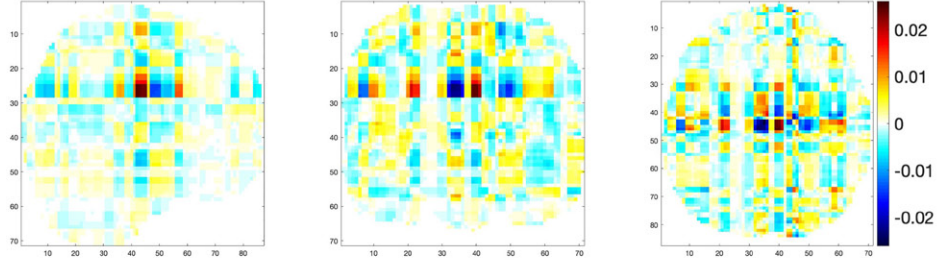
(c) TensorReg Lasso

(d) SHIV

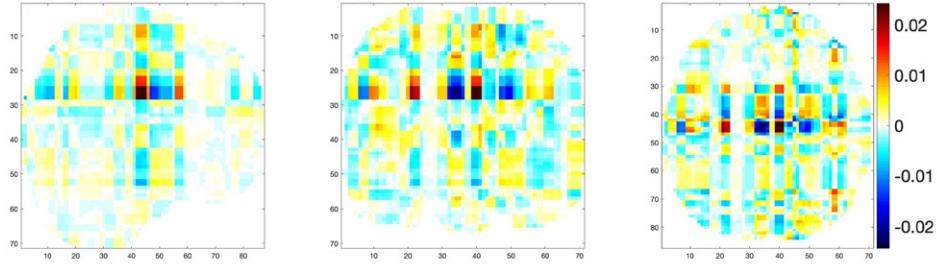
**Figure 3.** (Real data) Tensor coefficients estimated for PNC data using different methods. The first two rows, middle two rows, and last two rows are the sagittal section, the coronal section, and the horizontal section of a sample brain MRI, respectively. The color panels show the magnitude of all nonzero signals, and the black-white panels show only signals with magnitude that is seven standard deviations greater than the mean.



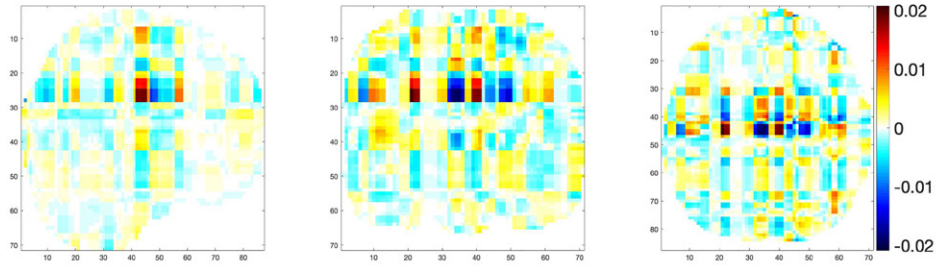
(a)  $\lambda = 0.5 \times 10^{-3}$ ,  $\text{BIC} = 7.5743 \times 10^3$



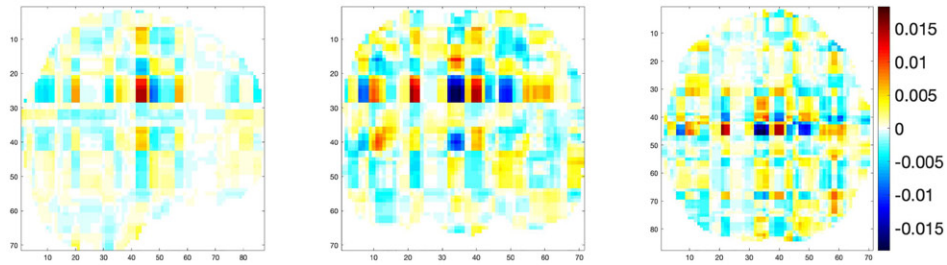
(b)  $\lambda = 0.75 \times 10^{-3}$ ,  $\text{BIC} = 7.4921 \times 10^3$ , selected



(c)  $\lambda = 1.0 \times 10^{-3}$ ,  $\text{BIC} = 7.5686 \times 10^3$



(d)  $\lambda = 1.25 \times 10^{-3}$ ,  $\text{BIC} = 7.7486 \times 10^3$



(e)  $\lambda = 1.5 \times 10^{-3}$ ,  $\text{BIC} = 8.1112 \times 10^3$

**Figure 4.** (Real data) Tensor coefficients estimated for PNC data using SHIV with different tuning parameter  $\lambda$ . From left to right, each column corresponds to the sagittal, the coronal, and the horizontal section of brain.



Our finding that a subregion across the cingulate cortex and the corpus callosum is associated with individuals' verbal reasoning ability has potentially important implications to the understanding of the underlying neurological mechanism behind this specific intellectual ability or even general intellectual abilities. It has been reported in the literature that cingulate cortex is associated with cognition (Sanfey et al. 2003; Leech and Sharp 2013). The anterior midcingulate cortex constitutes an information hub to reflect cognitive control (Shackman et al. 2011), and the middle cingulate cortex has extensive connections with the lateral prefrontal cortex (Stevens, Hurley, and Taber 2011), both of which are very close to, or slightly overlapped with, the subregion located by the proposed method. Meanwhile, subregions of corpus callosum are reported as being associated with dyslexia (Hynd et al. 1995; Von Plessen et al. 2002). In vocalization of emotions, the anterior cingulate cortex has been identified because of its connection with the Broca's area in the frontal lobes (Etkin, Egner, and Kalisch 2011). Therefore, our finding is not only novel, but also supported by the related literature. Moreover, both our finding and method are important to the future research in brain region identification. For example, high-order neuroimages, after preprocessing, can be directly input as covariates for scalable regression, while also have their structural and spatial information kept.

## Supplementary Materials

The supplement contains proofs for Propositions 1 and 2, and additional numerical studies with tables and figures.

## Acknowledgments

The authors thank the editor, associate editor, and two reviewers for insightful comments, and suggestions which improve the article significantly.

## Funding

This research was partially supported by grants R01 DA016750, R01 MH116527, and R01 HG010171 from the U.S. National Institutes of Health, and DMS1722544 from the National Science Foundation. Support for the collection of the PNC datasets was provided by NIH grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. All PNC participants were recruited through the Center for Applied Genomics at The Children's Hospital in Philadelphia. The PNC datasets were obtained from dbGaP at [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000607.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000607.v1.p1) through dbGaP accession: phs000607.v1.p1.

## ORCID

Long Feng  <http://orcid.org/0000-0002-2668-4805>

Xuan Bi  <http://orcid.org/0000-0002-4683-1411>

## References

- Baglio, F., Cabinio, M., Ricci, C., Baglio, G., Lipari, S., Griffanti, L., Preti, M. G., Nemni, R., Clerici, M., Zanette, M., and Blasi, V. (2014), "Abnormal Development of Sensory-Motor, Visual Temporal and Parahippocampal Cortex in Children With Learning Disabilities and Borderline Intellectual Functioning," *Frontiers in Human Neuroscience*, 8, 806. [144]
- Bajaj, S., Raikes, A., Smith, R., Dailey, N. S., Alkozei, A., Vanuk, J. R., and Killgore, W. D. (2018), "The Relationship Between General Intelligence and Cortical Structure in Healthy Individuals," *Neuroscience*, 388, 36–44. [145]
- Bathelt, J., Astle, D., Barnes, J., Raymond, F. L., and Baker, K. (2016), "Structural Brain Abnormalities in a Single Gene Disorder Associated With Epilepsy, Language Impairment and Intellectual Disability," *NeuroImage: Clinical*, 12, 655–665. [144,145]
- Burgaleta, M., Johnson, W., Waber, D. P., Colom, R., and Karama, S. (2014), "Cognitive Ability Changes and Dynamics of Cortical Thickness Development in Healthy Children and Adolescents," *Neuroimage*, 84, 810–819. [145]
- Daily, D. K., Ardinger, H. H., and Holmes, G. E. (2000), "Identification and Evaluation of Mental Retardation," *American Family Physician*, 61, 1059–1067. [144]
- Etkin, A., Egner, T., and Kalisch, R. (2011), "Emotional Processing in Anterior Cingulate and Medial Prefrontal Cortex," *Trends in Cognitive Sciences*, 15, 85–93. [154,157]
- Good, C. D., Johnsrude, I., Ashburner, J., Henson, R. N., Friston, K. J., and Frackowiak, R. S. (2001), "Cerebral Asymmetry and the Effects of Sex and Handedness on Brain Structure: A Voxel-Based Morphometric Analysis of 465 Normal Adult Human Brains," *Neuroimage*, 14, 685–700. [144]
- Gur, R. C., Alsop, D., Glahn, D., Petty, R., Swanson, C. L., Maldjian, J. A., Turetsky, B. I., Detre, J. A., Gee, J., and Gur, R. E. (2000), "An fMRI Study of Sex Differences in Regional Activation to a Verbal and a Spatial Task," *Brain and Language*, 74, 157–170. [144]
- Gur, R. C., Ragland, J. D., Moberg, P. J., Turner, T. H., Bilker, W. B., Kohler, C., Siegel, S. J., and Gur, R. E. (2001), "Computerized Neurocognitive Scanning: I. Methodology and Validation in Healthy People," *Neuropsychopharmacology*, 25, 766–776. [151]
- Gur, R. C., Richard, J., Calkins, M. E., Chiavacci, R., Hansen, J. A., Bilker, W. B., Loughhead, J., Connolly, J. J., Qiu, H., Mentch, F. D., and Abou-Sleiman, P. M. (2012), "Age Group and Sex Differences in Performance on a Computerized Neurocognitive Battery in Children Age 8–21," *Neuropsychology*, 26, 251. [145,150]
- Gur, R. C., Richard, J., Hughett, P., Calkins, M. E., Macy, L., Bilker, W. B., Brensing, C., and Gur, R. E. (2010), "A Cognitive Neuroscience-Based Computerized Battery for Efficient Measurement of Individual Differences: Standardization and Initial Construct Validation," *Journal of Neuroscience Methods*, 187, 254–262. [145,151]
- Hynd, G. W., Hall, J., Novey, E. S., Eliopoulos, D., Black, K., Gonzalez, J. J., Edmonds, J. E., Riccio, C., and Cohen, M. (1995), "Dyslexia and Corpus Callosum Morphology," *Archives of Neurology*, 52, 32–38. [154,157]
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002), "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images," *Neuroimage*, 17, 825–841. [150]
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012), "FSL," *Neuroimage*, 62, 782–790. [150]
- Jenkinson, M., and Smith, S. (2001), "A Global Optimisation Method for Robust Affine Registration of Brain Images," *Medical Image Analysis*, 5, 143–156. [150]
- Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [146]
- Kruskal, J. B. (1977), "Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, With Application to Arithmetic Complexity and Statistics," *Linear Algebra and its Applications*, 18, 95–138. [149]
- Lavin, C., Melis, C., Mikulan, E. P., Gelormini, C., Huepe, D., and Ibañez, A. (2013), "The Anterior Cingulate Cortex: An Integrative Hub for Human Socially-Driven Interactions," *Frontiers in Neuroscience*, 7, 64. [154]
- Leech, R., and Sharp, D. J. (2013), "The Role of the Posterior Cingulate Cortex in Cognition and Disease," *Brain*, 137, 12–32. [154,157]
- Lindquist, M. A. (2008), "The Statistical Analysis of fMRI Data," *Statistical Science*, 23, 439–464. [145]
- Liu, X., and Sidiropoulos, N. D. (2001), "Cramer-Rao Lower Bounds for Low-Rank Decomposition of Multidimensional Arrays," *IEEE Transactions on Signal Processing*, 49, 2074–2086. [149]
- Martin, A., and Chao, L. L. (2001), "Semantic Memory and the Brain: Structure and Processes," *Current Opinion in Neurobiology*, 11, 194–201. [144]

- Mehta, C. M., Gruen, J. R., and Zhang, H. (2017), "A Method for Integrating Neuroimaging Into Genetic Models of Learning Performance," *Genetic Epidemiology*, 41, 4–17. [154]
- 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. [145]
- Powell, M. J. (1964), "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," *The Computer Journal*, 7, 155–162. [150]
- Rudin, L. I., and Osher, S. (1994), "Total Variation Based Image Restoration With Free Local Constraints," in *Proceedings of 1st International Conference on Image Processing* (Vol. 1), pp. 31–35. [145,146]
- Rudin, L. I., Osher, S., and Fatemi, E. (1992), "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D: Nonlinear Phenomena*, 60, 259–268. [145,146]
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003), "The Neural Basis of Economic Decision-Making in the Ultimatum Game," *Science*, 300, 1755–1758. [154,157]
- Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., and Mentch, F. D. (2014), "Neuroimaging of the Philadelphia Neurodevelopmental Cohort," *Neuroimage*, 86, 544–553. [145,150]
- Schnack, H. G., Van Haren, N. E., Brouwer, R. M., Evans, A., Durston, S., Boomsma, D. I., Kahn, R. S., and Hulshoff Pol, H. E. (2015), "Changes in Thickness and Surface Area of the Human Cortex and Their Relationship With Intelligence," *Cerebral Cortex*, 25, 1608–1617. [145]
- Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., and Davidson, R. J. (2011), "The Integration of Negative Affect, Pain and Cognitive Control in the Cingulate Cortex," *Nature Reviews Neuroscience*, 12, 154–167. [157]
- Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., and Giedd, J. (2006), "Intellectual Ability and Cortical Development in Children and Adolescents," *Nature*, 440, 676–679. [144]
- Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., Foroud, T., Pankratz, N., Moore, J. H., Sloan, C. D., and Huentelman, M. J. (2010), "Whole Genome Association Study of Brain-Wide Imaging Phenotypes for Identifying Quantitative Trait Loci in MCI and AD: A Study of the ADNI Cohort," *Neuroimage*, 53, 1051–1063. [145]
- Sidiropoulos, N. D., and Bro, R. (2000), "On the Uniqueness of Multilinear Decomposition of N-Way Arrays," *Journal of Chemometrics*, 14, 229–239. [149]
- Smith, S. M. (2002), "Fast Robust Automated Brain Extraction," *Human Brain Mapping*, 17, 143–155. [150]
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., and Niazy, R. K. (2004), "Advances in Functional and Structural MR Image Analysis and Implementation as FSL," *Neuroimage*, 23, S208–S219. [150]
- Stevens, F. L., Hurley, R. A., and Taber, K. H. (2011), "Anterior Cingulate Cortex: Unique Role in Cognition and Emotion," *The Journal of Neuropsychiatry and Clinical Neurosciences*, 23, 121–125. [157]
- Ten Berge, J. M., and Sidiropoulos, N. D. (2002), "On Uniqueness in CANDECOMP/PARAFAC," *Psychometrika*, 67, 399–409. [149]
- Tibshirani, R. J. (2014), "Adaptive Piecewise Polynomial Estimation via Trend Filtering," *The Annals of Statistics*, 42, 285–323. [147]
- Tibshirani, R. J., and Taylor, J. (2011), "The Solution Path of the Generalized Lasso," *The Annals of Statistics*, 39, 1335–1371. [146,147,148]
- Von Plessen, K., Lundervold, A., Duta, N., Heiervang, E., Klauschen, F., Smievoll, A. I., Ersland, L., and Hugdahl, K. (2002), "Less Developed Corpus Callosum in Dyslexic Subjects—A Structural MRI Study," *Neuropsychologia*, 40, 1035–1044. [154,157]
- Vos, T., Barber, R. M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., Charlson, F., Davis, A., Degenhardt, L., Dicker, D., and Duan, L. (2015), "Global, Regional, and National Incidence, Prevalence, and Years Lived With Disability for 301 Acute and Chronic Diseases and Injuries in 188 Countries, 1990–2013: A Systematic Analysis for the Global Burden of Disease Study 2013," *The Lancet*, 386, 743–800. [144]
- Wang, X., Zhu, H., and Alzheimer's Disease Neuroimaging Initiative (2017), "Generalized Scalar-on-Image Regression Models via Total Variation," *Journal of the American Statistical Association*, 112, 1156–1168. [145,147]
- Zhang, Y., Brady, M., and Smith, S. (2001), "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm," *IEEE Transactions on Medical Imaging*, 20, 45–57. [150]
- 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. [145,146,148,149]
- Zhou, I. Y., Liang, Y.-X., Chan, R. W., Gao, P. P., Cheng, J. S., Hu, Y., So, K.-F., and Wu, E. X. (2014), "Brain Resting-State Functional MRI Connectivity: Morphological Foundation and Plasticity," *Neuroimage*, 84, 1–10. [145]