Human intelligence is usually measured by well-established psychometric tests through a series of problem solving. The recorded cognitive scores are continuous but usually heavy-tailed with potential outliers and violating the normality assumption. Meanwhile, magnetic resonance imaging (MRI) provides an unparalleled opportunity to study brain structures and cognitive ability. Motivated by association studies between MRI images and human intelligence, we propose a tensor quantile regression model, which is a general and robust alternative to the commonly used scalar-on-image linear regression. Moreover, we take into account rich spatial information of brain structures, incorporating low-rankness and piece-wise smoothness of imaging coefficients into a regularized regression framework. We formulate the optimization problem as a sequence of penalized quantile regressions with a generalized Lasso penalty based on tensor decomposition, and develop a computationally efficient alternating direction method of multipliers algorithm (ADMM) to estimate the model components. Extensive numerical studies are conducted to examine the empirical performance of the proposed method and its competitors. Finally, we apply the proposed method to a large-scale important dataset: the Human Connectome Project. We find that the tensor quantile regression can serve as a prognostic tool to assess future risk of cognitive impairment progression. More importantly, with the proposed method, we are able to identify the most activated brain subregions associated with quantiles of human intelligence. The prefrontal and anterior cingulate cortex are found to be mostly associated with lower and upper quantile of fluid intelligence. The insular cortex associated with median of fluid intelligence is a rarely reported region.

1. Introduction. Human intelligence depends on the cognitive abilities of problem-solving, which can be well measured by standardized psychometric tests that cover domains such as learning, reasoning, and memory (Deary, 2000). A detailed understanding of the brain mechanisms subserving this general cognitive ability could provide crucial individual and societal benefits. Modern neuroimaging techniques offer a unique opportunity to examine the relation between brain structure and function. In particular, magnetic resonance imaging (MRI) has become a prominent workhorse for in vivo quantitative measurements and the mapping of the human brain. See Toga and Thompson (2005) for a thorough discussion on how structural brain mapping could enhance our understanding of intelligence. However, identifying the neuroanatomical correlates of this relatively stable trait of intelligence has proved difficult, further hindered by the rapidly developing brains of adolescents (Shaw et al., 2006).

The Human Connectome Project (HCP) Consortium led by Washington University and the University of Minnesota (Van Essen et al., 2012, 2013) is a large-scale project spanning multiple types of data, including neuroimaging, genetics, as well as behavioral and cognitive assessment, with its overarching objective of revealing brain functions in healthy adults.
and understanding the diversity of human behaviors. Data were acquired from a population of 1200 healthy adults, in the age range of 22–35 years. Each participant undertook behavioral assessment through a well-validated battery, and multiple MR scanning sessions on the same equipment using the same protocol for acquiring neuroimaging data including structural MRI. We refer the readers to http://humanconnectome.org for detailed descriptions.

We are mostly interested in investigating how MRI images relate to the conditional distribution of fluid intelligence (cognitive scores) for the HCP cohort. This is challenging for several reasons. First, the cognitive scores often violate normality assumption and are heavy-tailed with potential outliers, which implies that the existing scalar-on-image linear regression (Zhou, Li and Zhu, 2013) may not be robust. Furthermore, the linear regression is restricted to the mean response rather than the whole response distribution. Second, the MRI images are ultrahigh dimensional but inherently contain spatial structural information, which suggests that the voxel-wise analysis and summarization-based methods may ignore the correlations, leading to the loss of statistical efficiency. For example, one may vectorize an MRI image of dimension $96 \times 120 \times 96$ as a vector of length $1.11 \times 10^6$, but applying the traditional linear quantile regression (Koenker and Bassett Jr, 1978) to the vectorized images destroys the spatial data structure and results in an ill-posed problem due to the ultrahigh dimensionality.

The literature on statistical analysis for neuroimaging has proliferated over the past decade, motivated by the increasingly collected imaging data. Zhou, Li and Zhu (2013) first introduced the tensor regression framework and applied it to the neuroimaging study. Zhou and Li (2014) proposed a regularized scalar-on-matrix linear regression for two-dimensional EEG data. Sun and Li (2017) introduced a linear regression model with a tensor response and scalar covariates. Wang and Zhu (2017) proposed to use a two-dimensional total variation for controlling smoothness of matrix coefficients in a linear regression. Li et al. (2018) introduced a tensor linear regression framework based on the Tucker decomposition. Recently, Feng, Bi and Zhang (2020) proposed an internal variation penalty for the tensor linear regression, which is an efficient substitute of total variation in high-order tensors.

Among the literature, most of them are motivated by studying brain responses to certain diseases, such as, Alzheimer’s disease (AD), and hence the methods are often concerned with binary responses through a generalized linear model. However, this is not the case in our study. Moreover, the HCP cohort consists of healthy young adults, which are different from populations with AD. It is also noteworthy that quantile regression has been widely applied to a variety of data types, such as longitudinal data (Wei et al., 2006) and genetic data (Briollais and Durrieu, 2014), but its application to neuroimaging has been rarely reported.

We propose a tensor quantile regression incorporating low-rank structure and piece-wise smoothness of the imaging coefficients to tackle the aforementioned challenges. Based on tensor decomposition, the estimation breaks into a sequence of regularized quantile regressions with a generalized Lasso penalty. The optimization is computationally challenging as the objective function is nonlinear in parameters, and the non-smooth loss function and penalty term further complicate the computation. Indeed, there is quite a bit of interest in researching sophisticated algorithms for sparse penalized quantile regression, see Yu and Lin (2017); Gu et al. (2018) and references therein, but our regularized quantile regression for tensor covariates is new. We develop a computationally efficient algorithm, which solves the generalized Lasso regularized quantile regression through a series of subproblems with analytical solutions. In contrast to existing methods, our method possesses several useful features. First, by employing dimension reduction, the tensor quantile regression can be used in all situations as a robust method to outliers and distribution assumption violations. Second, it offers us a complete analysis for associating brain images and cognitive scores beyond the mean response based regression. For example, exploiting extreme phenotypes may identify
more regions that are overlooked by the usual linear regression (Risch and Zhang, 1995). Finally, the tensor quantile regression can quantify uncertainty of the predictions using images and other covariates.

The remainder of the paper proceeds as follows: Section 2 introduces the proposed tensor quantile regression framework. Section 3 describes the estimation procedure and associated algorithms. In Section 4, we conduct extensive numerical studies to examine the empirical performance of the proposed method. In Section 5, we applied the proposed method to analyze the HCP dataset. Section 6 concludes this work with discussions.

2. Methodology.

2.1. Notations and Preliminaries. The observed data take the form \( \{(y_i, z_i, X_i) : i = 1, \ldots, n\} \), where \( y_i \) denotes the \( i \)th outcome (cognitive score), \( z_i \in \mathbb{R}^{p_0} \) is a vector of length \( p_0 \), which contains demographic covariates, \( X_i \in \mathbb{R}^{p_1 \times \cdots \times p_D} \) is a \( D \) dimensional tensor covariate, for example, an MRI image when \( D = 3 \), and \( n \) is the number of participants. In this study, we use the terms “array” and “tensor” interchangeably. Before introducing the model, we start by reviewing some useful notations and operations for tensors.

First, recall that \( \otimes \) represents the Kronecker product, and denote by \( \odot \) the column-wise Kronecker product. Let \( \text{vec}B \) represent the vectorization of a tensor \( B \in \mathbb{R}^{p_1 \times \cdots \times p_D} \), which stacks its elements into a column vector of length \( \prod_{d=1}^{D} p_d \) with its \( j \)th entry equal to the \((i_1, \cdots, i_D)\)th element of \( B \), where \( j = 1 + \sum_{d=1}^{D} (i_d - 1) \prod_{d' \neq d} p_{d'} \). Denote by \( B_{(d)} \) the mode-\( d \) matricization, which unfolds a tensor \( B \) into a \( p_d \times \prod_{d' \neq d} p_{d'} \) matrix with its \((i_d, j)\)th element mapping to the \((i_1, \cdots, i_D)\)th element of \( B \), where \( j = 1 + \sum_{d'' \neq d} (i_d' - 1) \prod_{d''' \neq d''} p_{d'''} \). Next, we define the inner product between two tensors as \( \langle A, B \rangle = \text{vec}B^\top \text{vec}A \), and the Frobenius norm of a tensor is defined as \( \|B\|_F = \sqrt{\langle B, B \rangle} \). Finally, we review the concept of the outer product. Let \( b_d \in \mathbb{R}^{p_d} \) be a vector of length \( p_d \), and the outer product of \( D \) vectors \( b_1 \odot \cdots \odot b_D \) is an array of dimension \( p_1 \times \cdots \times p_D \) with its \((i_1, \cdots, i_D)\)th element equal to \( \prod_{d=1}^{D} b_{d_{i_d}} \).

As a higher order analogue to matrix singular value decomposition, CANDECOMP/PARAFAC (CP) decomposition (Kolda and Bader, 2009) is commonly employed for dimension reduction. Assume that a \( D \)-way tensor \( B \in \mathbb{R}^{p_1 \times \cdots \times p_D} \) admits a rank \( R \) CP decomposition,

\[
B = \sum_{i=1}^{R} \beta_1^{(r)} \odot \cdots \odot \beta_D^{(r)},
\]

where \( \beta_d^{(r)} \in \mathbb{R}^{p_d} \) is a column vector corresponding to the \( d \)th mode and \( r \)th component, \( 1 \leq d \leq D, 1 \leq r \leq R \). We know that

\[
B_{(d)} = B_d \left( B_D \odot \cdots \odot B_{d+1} \odot B_{d-1} \odot \cdots \odot B_1 \right)^\top,
\]

where \( B_d \in \mathbb{R}^{p_d \times R} = \left[ \beta_d^{(1)}, \ldots, \beta_d^{(R)} \right] \). The CP decomposition of a tensor is not unique, so one may impose specific constraints to address the identifiability issue and use some sufficient and necessary conditions as practical tools to check the uniqueness, see Zhou, Li and Zhu (2013) and Feng, Bi and Zhang (2020).

2.2. Modeling Framework. Suppose we have a random triple \( (Y, z, X) \), where \( Y \) is a scalar response variable, \( z \) is a usual vector covariate, and \( X \) is a tensor covariate. The conditional distribution function of \( Y \) given the covariates is \( F_Y(y|z, X) = \mathbb{P}(Y \leq y|z, X) \). Given \( \tau \in (0, 1) \), the \( \tau \)th conditional quantile is defined as \( Q_\tau(Y|z, X) = \inf \{ y : F_Y(y|z, X) \geq \tau \} \).
We consider the conditional quantile taking a linear form to model the relationship between the \(\tau\)th quantile of the response and the tensor covariate as well as scalar covariates,

\[
Q_{\tau}(Y|\mathbf{z}, \mathbf{X}) = \mathbf{z}^\top \gamma_{\tau} + < \mathbf{X}, \mathbf{B}_{\tau} > ,
\]

where \(\gamma_{\tau} = (\gamma_{\tau 1}, \cdots, \gamma_{\tau p})^\top\) is the coefficient vector corresponding to the demographic covariates, and \(\mathbf{B}_{\tau}\) is the tensor coefficient quantifying the effect of tensor covariate \(\mathbf{X}\) at the quantile level \(\tau\). Model (1) provides a general and robust alternative to the tensor linear regression proposed in Zhou, Li and Zhu (2013).

As in Koenker (2005), a direct approach for the scalar-on-image quantile regression is minimizing the following loss function for the \(\tau\)th quantile,

\[
\ell(\gamma_{\tau}, \mathbf{B}_{\tau}) = \sum_{i=1}^{n} \rho_{\tau}(y_i - \mathbf{z}_i^\top \gamma_{\tau} - < \mathbf{X}_i, \mathbf{B}_{\tau} > ),
\]

where \(\rho_{\tau}(x) = x\{\tau - 1(x \leq 0)\}\) is the so-called check function.

Alternatively, we may connect the tensor quantile regression with the generalized tensor linear regression (Zhou, Li and Zhu, 2013) by treating the indicator of \(Y \leq y\) as an auxiliary binary response and modeling the empirical distribution function. This approach has been deployed in quantile regression with scalar covariates (Meinshausen, 2006). Specifically, given a fixed \(y\),

\[
F_{Y}(y|\mathbf{z}, \mathbf{X}) = \mathbb{E}\{\mathbf{1}(Y \leq y)|\mathbf{z}, \mathbf{X}\} = g^{-1}\left\{ \mathbf{z}^\top \gamma_y + < \mathbf{X}, \mathbf{B}_y > \right\},
\]

where \(g(\cdot)\) is a known, monotone link function, namely, the logit link function defined as \(g^{-1}(x) = \exp(x)/(1 + \exp(x))\). We can estimate \(\gamma_y\) and \(\mathbf{B}_y\), and then obtain all the quantiles simultaneously by inverting the fitted distribution function.

Although the indirect approach (3) is straightforward to implement, in this work we focus on the direct approach (2) since it enables us to study the association between tensor covariates and specific quantiles other than solely predicting the response. To improve interpretability of the model and detect the potential brain subregions highly relevant to fluid intelligence, we propose a regularized framework which imposes sparsity structures via a class of penalty terms on the tensor coefficient \(\mathbf{B}_{\tau}\),

\[
\sum_{i=1}^{n} \rho_{\tau}(y_i - \mathbf{z}_i^\top \gamma_{\tau} - < \mathbf{X}_i, \mathbf{B}_{\tau} > ) + P_\lambda(\mathbf{B}_{\tau}),
\]

where \(P_\lambda(\mathbf{B}_{\tau})\) represents penalty terms depending on \(\lambda\), which is a non-negative tuning parameter balancing model fitting and complexity. When \(\lambda = 0\), the regularized tensor quantile regression (4) reduces to the tensor quantile regression (2). Specifically, we consider several representative penalties: the Lasso \(\lambda \|\mathbf{B}_{\tau}\|_1\) as in Zhou, Li and Zhu (2013), which penalizes each marginal vector of mode \(d\) with an \(l_1\) norm penalty and hence it leads to an overall sparsity of the tensor; the internal variation \(\lambda \|\mathbf{B}_{\tau}\|_{IV}\) proposed in Feng, Bi and Zhang (2020), which encourages piece-wise constant structures of the tensor coefficient; and a fusion penalty \(\lambda \|\mathbf{B}_{\tau}\|_{fusion}\) imposing a generalized Lasso along each mode, which also ensures piece-wise smoothness of the tensor coefficient effects and enjoys more flexibility. Piece-wise smoothness is a sensible choice for characterizing the brain regional effects since voxels in close proximity tend to have similar coefficients, albeit they might be at or across the boundary of brain anatomical regions. In this sense, this structure is relevant to the capability of discovering subregions which may not necessarily belong to a single anatomical region, or interplay among multiple brain regions. The term \(P_\lambda(\mathbf{B}_{\tau})\) is general as we will show that all the three penalties can be unified in a single formulation and solved by an efficient algorithm.
3. Estimation.

3.1. Regularized Quantile Regression with Generalized Lasso Penalty. Estimation of $B_τ$ in problem (2) can follow the block relaxation algorithm as in Zhou, Li and Zhu (2013). We now briefly describe the procedure. Recall that the tensor coefficient $B_τ$ is assumed to follow a CP decomposition with $R$ components. Minimizing the loss function (2) is equivalent to minimizing

$$\sum_{i=1}^{n} \rho_r(y_i - z_i^T \gamma_τ - <X_i(d), B_{τd} (B_{τD} \odot \cdots \odot B_{τ,d+1} \odot B_{τ,d-1} \odot \cdots \odot B_{τ1})^T >)$$

$$= \sum_{i=1}^{n} \rho_r(y_i - z_i^T \gamma_τ - <B_{τd}, X_i(d) (B_{τD} \odot \cdots \odot B_{τ,d+1} \odot B_{τ,d-1} \odot \cdots \odot B_{τ1}) >),$$

which implies that $\beta_{τd} \in \mathbb{R}^{R_{τd}} = \text{vec}B_{τd} = (\beta_{τ1}^T, \cdots, \beta_{τR}^T)^T$ can be updated iteratively if other $B_{τ,d' \neq d}$ are fixed. This turns out to be a sequence of traditional quantile regressions via practical algorithms, for example, the interior point algorithm, and $γ_τ$ can be estimated separately once each mode of $B_τ$ is updated.

### Table 1

| Penalty terms incorporated in the regularized tensor quantile regression (4). We assume the tensor coefficient $B_τ$ follows a CP decomposition with $R$ components. All the penalties can be generalized as $\lambda|D_{τd}\beta_{τd}|$ for mode $d$ with varying weights matrix $D_{τd}$. |
|---|---|---|---|
| $P_λ(B_τ)$ | Structure | Formulation | $D_{τd}\beta_{τd}$ |
| $\lambda|B_τ|$ | sparse | $\sum_{r=1}^{R} \sum_{d=1}^{D} |\beta_{τd}^r|$ | $I_{R} \otimes I_d$ |
| $\lambda||B_τ||_1$ | piece-wise constant | $\sum_{r=1}^{R} \prod_{d=1}^{D} ||\beta_{τd}^r||_1$ | $\sum_{r=1}^{R} |\beta_{τd}^r| \circ \sum_{r=1}^{R} \Lambda_{τd}^{(r)} |R_{d}^{(r)}| \beta_{τd}^r$ | $I_{R} \otimes \sum_{r=1}^{R} |\beta_{τd}^r| \circ \sum_{r=1}^{R} \Lambda_{τd}^{(r)} |R_{d}^{(r)}|$ |
| $\lambda||B_τ||_1$ | piece-wise smooth | $\sum_{r=1}^{R} \sum_{d=1}^{D} |\beta_{τd}^m|$ | $\sum_{r=1}^{R} |\beta_{τd}^m| \circ \sum_{r=1}^{R} \Lambda_{τd}^{(r)} |R_{d}^{(r)}| \beta_{τd}^r$ | $I_{R} \otimes \sum_{r=1}^{R} |\beta_{τd}^m| \circ \sum_{r=1}^{R} \Lambda_{τd}^{(r)} |R_{d}^{(r)}|$ |

Now we consider the optimization problem for the regularized formulation (4). Note that the three penalty terms can be unified by a generalized Lasso penalty $\lambda|D_{τd}\beta_{τd}|$ for each mode $d$, where $D_{τd}$ is a weights matrix. Then, we estimate each $\beta_{τd}$ iteratively through a sequence of penalized quantile regressions,

$$\sum_{i=1}^{n} \rho_r(y_i - z_i^T \gamma_τ - x_{τd}^T \beta_{τd}) + \lambda|D_{τd}\beta_{τd}|,$$

Table 1 summarizes the three penalties and provides details for computation. The fourth column of Table 1 presents the specific formulations along mode $d$, and the unified penalty only varies in $D_{τd}$, shown in the last column of Table 1. For the Lasso, let $I_R \in \mathbb{R}^{R \times R}$ and $I_d \in \mathbb{R}^{p_d \times p_d}$ be two identity matrices, then $D_{τd} = D_d = I_R \otimes I_d$. For the internal variation, denote by $\|b\|_1 = \sum_{j=2}^{p} |b_j - b_{j-1}|$ a one-dimensional total variation penalty (Rudin, Osher and Fatemi, 1992) for any $b \in \mathbb{R}^p$, then we have $\|\beta_{τd}^r\|_1 = |R_{d}^{(r)}| \beta_{τd}^r|$, where $R_{d}^{(r)} \in \mathbb{R}^{(p_d-1) \times p_d}$ is a first-order differencing matrix

$$R_{d}^{(1)} = \begin{pmatrix}
1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & -1
\end{pmatrix}.$$
and we write $\prod_{d \neq d'} \|eta_{rd}^{(r)}\|_{TV}$ as $A_{rd}^{(r)}$, it leads to $D_{rd} = \text{diag}(A_{rd}^{(1)}, \cdots, A_{rd}^{(R)}) \otimes R_{d}^{(1)}$. For the fusion penalty, one may consider imposing a piece-wise polynomial structure of degree $m - 1$ to $\beta_{rd}^{(r)}$ through the $m$th-order differencing matrix $R_{d}^{(m)}$, which is defined recursively by left multiplying a first-order differencing matrix. It can be shown that $D_{rd} = D_{d} = I_{R} \otimes R_{d}^{(m)}$ in this setup. When $m = 2$, the penalty enforces a piece-wise linear structure, commonly used in trend filtering (Kim et al., 2009; Brantley, Guinness and Chi, 2020).

We present the overall algorithm for optimizing the regularized tensor quantile regression (4) as in Algorithm 1. When we consider the basic model (2), Step 4 can be solved by the interior point algorithm as in Step 6 of Algorithm 1.

Algorithm 1: Block relaxation algorithm for minimizing (4).

1. Initialize $\gamma_{\tau}^{(0)}, \beta_{rd}^{(0)} \in \mathbb{R}^{p_d R}$ for $d = 1, \cdots, D$;
2. repeat
3. for $d = 1, \cdots, D$ do
4. $\beta_{rd}^{(t+1)} = \arg \min_{\beta_{rd}} \left\{ \sum_{i=1}^{n} \rho_{\tau}(y_i - z_i^{\top} \gamma_{\tau}^{(t)} - x_{i, rd} \beta_{rd}) + \lambda |D_{rd}^{(t)} \beta_{rd}| \right\}$
5. end
6. $\gamma_{\tau}^{(t+1)} = \arg \min_{\gamma_{\tau}} \left\{ \sum_{i=1}^{n} \rho_{\tau}(y_i - z_i^{\top} \gamma_{\tau} - \langle \mathcal{X}_i, B_{\tau}^{(t+1)} \rangle) \right\}$
7. until $\frac{|| (\gamma_{\tau}^{(t+1)}, B_{\tau}^{(t+1)}) - (\gamma_{\tau}^{(t)}, B_{\tau}^{(t)}) ||}{|| (\gamma_{\tau}^{(t+1)}, B_{\tau}^{(t+1)}) ||} < 10^{-4}$;

3.2. ADMM Algorithm. We now drop the subscripts $d$ and $\tau$ for brevity, and let $\tilde{y}_i = y_i - z_i^{\top} \gamma$. The penalized quantile regression (5) can be rewritten as

$$
\sum_{i=1}^{n} \rho_{\tau}(\tilde{y}_i - x_i^{\top} \beta) + \lambda |D| \beta.
$$

Solving this problem is challenging since both the loss function and the penalty term are non-smooth. Furthermore, this problem may become a computational bottleneck of the overall block relaxation algorithm as it is replicated for each mode in the main loop. Therefore, we derive a computationally efficient alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011) for minimizing the quantile regression with a generalized Lasso penalty.

First, it is natural to reformulate optimization problem (6) as

$$
\sum_{i=1}^{n} \rho_{\tau}(w_i) + \lambda |D| \beta, \quad \text{subject to} \quad w = \tilde{y} - X \beta,
$$

where $w = (w_1, \cdots, w_n)^{\top}, \tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_n)^{\top}, X = (x_1, \cdots, x_n)^{\top}$. Then, we can solve the augmented Lagrangian as in Boyd et al. (2011). However, there is no analytical solution for solving the generalized Lasso, and hence an embedded iterative solver for $\beta$ is needed. This could be either a coordinate decent algorithm as in Tibshirani and Taylor (2011) or another ADMM algorithm, which may decelerate the whole process. We defer the technical details of this algorithm to the Supplementary Material (Li and Zhang, 2021).

Alternatively, we propose to solve the following problem with one extra constraint,

$$
\sum_{i=1}^{n} \rho_{\tau}(w_i) + \lambda |z|, \quad \text{subject to} \quad w = \tilde{y} - X \beta, \quad z = D \beta,
$$
where \( z = (z_1, \cdots, z_q)^T \). The constrained form is equivalent to

\[
\sum_{i=1}^{n} \rho_{\tau}(w_i) + \lambda |z| + \langle \theta_1, \tilde{y} - w - X\beta \rangle + \frac{\alpha}{2} \|\tilde{y} - w - X\beta\|^2 \\
+ \langle \theta_2, D\beta - z \rangle + \frac{\alpha}{2} \|D\beta - z\|^2,
\]

(8)

where \( \| \cdot \| \) is the \( \ell_2 \) norm, \( \theta_1 \) and \( \theta_2 \) are Lagrangian multipliers, \( \alpha > 0 \) is a tuning parameter. Although we introduce new variable \( z \) and its dual variable \( \theta_2 \), the benefit is that now we are able to obtain analytical solutions for \( z \) and \( \beta \). Indeed, we shall see that the algorithm translates the difficult problem (6) into a series of simple subproblems, each with a closed-form solution and thus is computationally easy.

Next, we estimate each model components in an iterative manner since the parameters are completely decoupled in (8). The algorithm consists of the iterations, which correspond to five subproblems:

\[
z^{(k+1)} = \arg \min_z \left\{ \lambda |z| + \frac{\alpha}{2} \|z - D\beta^{(k)} - \frac{1}{\alpha} \theta_2^{(k)}\|^2 \right\},
\]

(9)

\[
\beta^{(k+1)} = \arg \min_{\beta} \left\{ \frac{\alpha}{2} \|\tilde{y} - w^{(k)} - X\beta\|^2 + \frac{\alpha}{2} \|D\beta - z^{(k+1)}\|^2 - (\theta_1^{(k)})^T X - \theta_2^{(k)^T} D\beta \right\},
\]

(10)

\[
w^{(k+1)} = \arg \min_w \left\{ \sum_{i=1}^{n} \rho_{\tau}(w_i) - \langle \theta_1^{(k)}, w \rangle + \frac{\alpha}{2} \|\tilde{y} - X\beta^{(k+1)} - w\|^2 \right\},
\]

(11)

\[
\theta_1^{(k+1)} = \theta_1^{(k)} - \alpha (w^{(k+1)} + X\beta^{(k+1)} - \tilde{y}),
\]

\[
\theta_2^{(k+1)} = \theta_2^{(k)} - \alpha (z^{(k+1)} - D\beta^{(k+1)}).
\]

To update \( z \), a closed-form solution to (9) is

\[
\tilde{z} = \text{Soft}_{\lambda/\alpha} \left( D\beta + \frac{1}{\alpha} \theta_2 \right),
\]

where the soft-thresholding operator is defined as

\[
\text{Soft}_{\lambda}(x) = \begin{cases} 
  x - \lambda, & \text{if } x > \lambda, \\
  0, & \text{if } -\lambda \leq x \leq \lambda, \\
  x + \lambda, & \text{if } x < -\lambda.
\end{cases}
\]

To update \( \beta \), problem (10) only involves linear and quadratic forms of \( \beta \), it leads to

\[
\tilde{\beta} = \left( X^T X + D^T D \right)^{-1} \left\{ \frac{1}{\alpha} X^T \theta_1 - D^T \theta_2 \right\} + X^T (\tilde{y} - w) + D^T z.
\]

To update \( w \), note that \( \rho_{\tau}(w) = w \{ \tau - 1(w < 0) \} = \frac{1}{2} |w| + (\tau - 1/2) w \), then it can be shown that minimizing (11) is equivalent to minimizing

\[
\frac{1}{2} |w| + \frac{\alpha}{2} \|w - \left\{ \tilde{y} - X\beta + \frac{1}{\alpha} \theta_1 - \frac{1}{\alpha} \left( \tau - \frac{1}{2} \right) \right\}\|^2.
\]

Similarly, we have

\[
\tilde{w} = \text{Soft}_{1/2\alpha} \left( \tilde{y} - X\beta + \frac{1}{\alpha} \theta_1 - \frac{1}{\alpha} \left( \tau - \frac{1}{2} \right) \right).
\]
In our numerical experiments, the proposed algorithm greatly alleviates the computational burden, and shows significant speedup compared with the previous algorithm of one constraint. We will show its scalability in Section 5, which is found to be comparable to the state-of-the-art optimization techniques exploited by the regularized tensor linear regression.

The ADMM algorithm has the following convergence guarantee.

**Proposition 1.** As $k \to \infty$, $\sum_{i=1}^{n} \rho_i (w_i^{(k)}) + \lambda |z^{(k)}|$ converges to the minimum objective value of problem (7): $\| \hat{y} - X\beta^{(k)} - w^{(k)} \|$ and $\| D\beta^{(k)} - z^{(k)} \|$ converge to zero.

Proposition 1 holds as a direct application of the convergence results in Boyd et al. (2011). Following Boyd et al. (2011), we define the primal and dual residues at the $k$th iteration as

$$s^{(k)}_{\text{primal}} = \sqrt{\| \hat{y} - X\beta^{(k)} - w^{(k)} \|^2 + \| D\beta^{(k)} - z^{(k)} \|^2},$$

$$s^{(k)}_{\text{dual}} = \rho \sqrt{\| X^\top (w^{(k)} - w^{(k-1)}) \|^2 + \| D^\top (z^{(k)} - z^{(k-1)}) \|^2}.$$

Then, a reasonable stopping criteria is that the primal and dual residuals are less than some thresholds,

$$s^{(k)}_{\text{primal}} \leq \sqrt{n} \epsilon_1 + \epsilon_2 \max \left\{ \| \hat{y} \|, \| X\beta^{(k)} \|, \| w^{(k)} \|, \| D\beta^{(k)} \|, \| z^{(k)} \| \right\},$$

$$s^{(k)}_{\text{dual}} \leq \sqrt{n} \epsilon_1 + \epsilon_2 \sqrt{\| X^\top \Theta_1^{(k)} \|^2 + \| D^\top \Theta_2^{(k)} \|^2}.$$

We set $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-2}$ in our numerical studies and terminate the iterations when the stopping criteria are satisfied or the maximum number of iterations, such as $10^3$, are reached.

### 3.3. Model Selection.

In practice, we select the tuning parameters $R$ and $\lambda$ according to a Schwarz-type information criterion (SIC) as in Koenker, Ng and Portnoy (1994); He, Ng and Portnoy (1998), or the generalized approximate cross-validation criterion (GACV) as in Yuan (2006),

$$\text{SIC}(R, \lambda) = \log \left\{ \frac{\ell(\hat{\tau}_\tau, \hat{B}_\tau)}{n} \right\} + \frac{\log(n)}{2n} \cdot \text{df}(R, \lambda),$$

$$\text{GACV}(R, \lambda) = \frac{\ell(\hat{\tau}_\tau, \hat{B}_\tau)}{n - \text{df}(R, \lambda)},$$

where $\text{df}(R, \lambda)$ is the degrees of freedom. One sensible choice might be calculating $\text{df}(R, \lambda)$ as in Feng, Bi and Zhang (2020) and Brantley, Guinness and Chi (2020), which is the sum of the number of nonzero elements in $D_d \beta_{\tau_d}$ for $d = 1, \ldots, D$, and it is found to work well. For each combination of $R$ and $\lambda$, we fit the model to obtain the estimates $\hat{\tau}_\tau$ and $\hat{B}_\tau$, and evaluate the SIC or GACV based on $\ell(\hat{\tau}_\tau, \hat{B}_\tau)$ and $\text{df}(R, \lambda)$. We use SIC in our numerical studies since the selection of GACV is found to be close to that of SIC.

### 4. Simulation Studies.

In this section, we carry out simulation studies to examine the empirical performance of the proposed methods. Specifically, we compare tensor quantile regression and its regularized form, denoted by “TQR” and “RTQR” respectively, throughout the paper, with several other competitors: tensor linear regression as proposed in Zhou, Li and Zhu (2013), denoted by “TR”; the regularized tensor linear regression with the internal variation penalty as proposed in Feng, Bi and Zhang (2020), denoted by “SHIV”; the voxel-wise quantile regression, which conducts quantile regression for the response against each
voxel that can be directly adopted by practitioners, denoted by “VoxelQR.” We use SHIV as a representative of the regularized tensor linear regression as it shows the best overall performance among others (Feng, Bi and Zhang, 2020). Similarly, we focus on RTQR with the fusion penalty using $R^{(1)}_d$ in our simulation studies and real data application, and RTQR with the Lasso and internal variation penalty also show competitive performance.

Following Zhou, Li and Zhu (2013) and Feng, Bi and Zhang (2020), we generate responses according to

$$y = z^\top \gamma + \langle X', B \rangle + \epsilon,$$

where $z$ is a covariate vector containing five elements with each one sampled independently from a standard normal distribution, and $\gamma$ is the corresponding parameter vector with all elements equal to one. Let $X'$ be a matrix or tensor covariate according to the simulation setup that will be introduced later, and its entries are also standard normals. The matrix/tensor parameter $B$ is binary with true signal region equal to one and the rest zero. The random errors $\epsilon$ are sampled independently from different distributions. Several factors are mainly considered in the numerical experiments: first, to assess the robustness of tensor quantile regression, we generate random errors from standard normal, standard Cauchy, and t distribution with 3 degrees of freedom, respectively; second, we vary the shape of the tensor parameter and adopt the 2D and 3D setup as in Zhou, Li and Zhu (2013) and Feng, Bi and Zhang (2020), respectively; third, three levels of sample size are used: $n = 600, 800, 1200$. In addition, 200 new subjects are generated as the testing set for evaluating the prediction performance. The simulations are replicated for 100 times.

Estimation accuracy is evaluated by squared error $SE = \|B - \hat{B}\|_F^2$, or its root mean squared error (RMSE). For a fair comparison, we use the median estimate $\hat{B}_{0.5}$ of the quantile regression for the assessment. Mean absolute error (MAE) of the corresponding quantiles $\sum_{i=1}^{200} |Q_{i, \tau} - \hat{Q}_{i, \tau}|/200$ are computed as the evaluation criterion for prediction, where $\hat{Q}_{i, \tau}$ represents the $\tau$th predicted quantile for subject $i$. Note that for those tensor regressions, we only calculate MAE for the discrepancy between medians and predicted responses.

4.1. 2D Simulations. To illustrate the proof-of-concept, we start by comparing TR and TQR, and will systematically examine the performance of the advanced methods in Section 4.2. We generate $X'$ as a $64 \times 64$ two-dimensional matrix, and use the matrix signals as in Zhou, Li and Zhu (2013). Following their setup, we simulate the data by setting the ranks of the signal tensors as 1 for “square,” 2 for “T” and “cross,” and 3 for the rest, respectively. See Figure S.1 in the Supplementary Material (Li and Zhang, 2021) for the detailed results of selecting ranks according to the model selection criteria introduced in Section 3.3. We fit TQR for $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Figure 1 and Figure S.2 in the Supplementary Material (Li and Zhang, 2021) show that TQR is able to identify the signal regions in all cases, but TR can hardly recognize any distinguishable regions in the Cauchy case. This implies that TQR outperforms TR remarkably in terms of estimating $B$ when the random errors follow a Cauchy distribution. Furthermore, the performance of TQR is comparable to that of TR when the errors are normally distributed where linear regression is the most efficient method (Koenker and Bassett Jr, 1978). The scenario of t distribution shows similar patterns as the Cauchy case, and thus are omitted for simplicity. Figure 2 presents the results of predicting quantiles. Again, while TR performs best when the errors are Gaussian, TQR is close to TR; when the errors are t-distributed with heavy tails, the performance of TQR for $\tau = 0.5$ is significantly superior to that of TR; when the errors are Cauchy-distributed with heavier tails, the performance of TR drastically degenerates, and TQR has much smaller MAEs than those of TR at all levels. We also vary the noise level in the experiments and the patterns are consistent.
Fig 1: True signals and estimates of matrix parameters by TR and TQR. The sample size is 800 and the random errors follow a Cauchy distribution. Tensor regressions are fitted with rank $r$. Each estimate of TR and TQR is a median of 100 replications. RMSE is calculated based on $\|B - \hat{B}\|_F^2$. 
4.2. 3D Simulations. Here we generate $\mathcal{X}$ as a $30 \times 30 \times 30$ tensor and use three-dimensional signals of various shapes as in Feng, Bi and Zhang (2020). We fix the ranks as 1, 2, 3, and 4 for “one brick,” “two bricks,” “cross,” and “pyramid” respectively as those low-rank tensors can well approximate the truth. We then choose tuning parameters of the regularized regressions over a grid according to SIC. Three quantile levels $\tau = \{0.25, 0.5, 0.75\}$ are considered, and we focus on the normal and Cauchy cases.

Table 2 with various model conditions provides detailed results, and we make the following key observations: first, VoxelQR has the largest SEs in all cases due to its marginal nature that fails to respect the structural information of the tensor; second, it is not surprising that SHIV performs best for the normal case, but its performance is compromised by Cauchy errors, which makes its SEs the second worst in this scenario; third, when the errors are non-Gaussian, TQR outperforms SHIV, and RTQR further improves TQR and achieves the best overall performance. When the errors are Gaussian, TQR and RTQR are comparable to SHIV; fourth, all the methods improve as the sample size increases, and the benefits of a larger sample size are the most profound for TQR and RTQR. Last but not the least, all the methods yield smaller SEs when the true signal poses a simpler structure. Furthermore, Figure 3 and Figure S.3 in the Supplementary Material (Li and Zhang, 2021) visualize the estimated tensor parameters of various methods, except for VoxelQR as it can never recover the true signals. For the Cauchy case, the regions identified by RTQR are the closest to the truth; TQR can also recover the signals but presents some biases for the boundary regions; SHIV shows the largest deviation from the true signals. For the Gaussian case, all the methods recognize the signal regions satisfactorily.

Table 3 presents the performance of various methods in terms of prediction. The pattern is consistent with that of estimation accuracy: TQR and RTQR are close to TR when the Gaussian assumption is not violated, but they are much more robust and accurate than TR when the errors are non-Gaussian; RTQR uniformly improves TQR; the prediction accuracy is reasonable with an increasing sample size. Note that VoxelQR cannot be used for predicting quantiles directly as each voxel may result in a separate prediction.
5. Application to Neuroimaging Study. For the HCP data, we employ the 1200-subject release in our study and investigate the relationship between brain regions and fluid intelligence, which is measured using an abbreviated version of Raven’s Progressive Matrices (Bilker et al., 2012). Participants are presented with patterns consisted of $2 \times 2$, $3 \times 3$ or $1 \times 5$ arrangements of squares, with one of the squares missing. The participant must choose the correct response that best fits the missing square on the pattern from five choices. The battery has 24 items and 3 bonus items in total and ordered by increasing difficulty. However, the task discontinues if the participant fails in 5 consecutive times. The number of correct responses ranging from 0 to 27 is considered as the response of interest.

We use the T1-weighted (T1w) MRI data as neuroimaging covariates, which were acquired on the HCP’s custom 3T Siemens Skyra MRI scanner using a 32-channel head coil (Uğurbil et al., 2013). The T1w images were acquired using the 3D MPRAGE sequence (Mugler and Brookeman, 1990) with 0.7 mm isotropic resolution. The images have been fully prepared by the HCP minimal preprocessing pipelines (Glasser et al., 2013), including three structural pipelines: PreFreeSurfer, FreeSurfer, and PostFreeSurfer. The major preprocessing steps of PreFreeSurfer included gradient distortion correction, co-registration and averaging of T1w runs, anterior commissure–posterior commissure (ACPC) registration, brain extraction, field map distortion correction, and bias field correction. Structural MRI scans were acquired in arbitrary units, making intensity-based analysis impossible without intensity normalization; see, for example, Shinohara et al. (2014) and Reinhold et al. (2019), for some discussions of

![Fig 3](image-url)

Fig 3: True signals and estimates of tensor parameters. The sample size is 800 and the random errors follow a Cauchy distribution. Each estimate is a median of 100 replications.
Table 2

<table>
<thead>
<tr>
<th>Case</th>
<th>Signal shape</th>
<th>n</th>
<th>VoxelQR</th>
<th>SHIV</th>
<th>TQR</th>
<th>RTQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>One brick</td>
<td>600</td>
<td>8012.45</td>
<td>(3383.55)</td>
<td>0.04</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>7424.02</td>
<td>(2460.4)</td>
<td>0.03</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>4505.71</td>
<td>(1160.86)</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Two bricks</td>
<td>600</td>
<td>18414.06</td>
<td>(6236.69)</td>
<td>0.1</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>13259.56</td>
<td>(1889.25)</td>
<td>0.06</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>9097</td>
<td>(3595.28)</td>
<td>0.03</td>
<td>0.25</td>
</tr>
<tr>
<td>Normal</td>
<td>Cross</td>
<td>600</td>
<td>32811.2</td>
<td>(9651.18)</td>
<td>0.28</td>
<td>2.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>24199.91</td>
<td>(7761.03)</td>
<td>0.2</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>15845.59</td>
<td>(3312.15)</td>
<td>0.15</td>
<td>0.43</td>
</tr>
<tr>
<td>Normal</td>
<td>Pyramid</td>
<td>600</td>
<td>6549.86</td>
<td>(2406.39)</td>
<td>30.51</td>
<td>72.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>4281.67</td>
<td>(1434.82)</td>
<td>1.39</td>
<td>3.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>2715.44</td>
<td>(547.28)</td>
<td>0.36</td>
<td>0.6</td>
</tr>
<tr>
<td>Normal</td>
<td>Cauchy</td>
<td>600</td>
<td>9551.17</td>
<td>(2933.53)</td>
<td>281.09</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>7827.87</td>
<td>(1483.72)</td>
<td>221.95</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>4777.19</td>
<td>(1243.14)</td>
<td>60.37</td>
<td>0.23</td>
</tr>
<tr>
<td>Normal</td>
<td>Pyramid</td>
<td>600</td>
<td>21958.89</td>
<td>(7069.43)</td>
<td>665.61</td>
<td>406.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>15032.26</td>
<td>(4018.48)</td>
<td>373.27</td>
<td>211.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>9252.29</td>
<td>(2202.75)</td>
<td>133.91</td>
<td>0.67</td>
</tr>
<tr>
<td>Normal</td>
<td>Cauchy</td>
<td>600</td>
<td>34545.98</td>
<td>(5132.59)</td>
<td>1093.93</td>
<td>868.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>26516.8</td>
<td>(7131.53)</td>
<td>595.3</td>
<td>266.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>16714.2</td>
<td>(3701.12)</td>
<td>760.25</td>
<td>171.41</td>
</tr>
<tr>
<td>Normal</td>
<td>Cauchy</td>
<td>600</td>
<td>3254.74</td>
<td>(1472.36)</td>
<td>330.75</td>
<td>20.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>5268.33</td>
<td>(1550.45)</td>
<td>671.09</td>
<td>137.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>7718.19</td>
<td>(1483.32)</td>
<td>910.07</td>
<td>202.53</td>
</tr>
<tr>
<td>Normal</td>
<td>Pyramid</td>
<td>600</td>
<td>5268.33</td>
<td>(1550.45)</td>
<td>671.09</td>
<td>137.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>3254.74</td>
<td>(1472.36)</td>
<td>330.75</td>
<td>20.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1200</td>
<td>7718.19</td>
<td>(1483.32)</td>
<td>910.07</td>
<td>202.53</td>
</tr>
</tbody>
</table>

the importance of intensity normalization in structural images. In our analysis, we use images further prepared by FreeSurfer and PostFreeSurfer pipelines, which contained the following steps: segmentation of predefined structures, intensity normalization, reconstruction of white
<table>
<thead>
<tr>
<th>Case</th>
<th>Signal shape</th>
<th>$\tau$</th>
<th>TQR</th>
<th>RTQR</th>
<th>TQR</th>
<th>SHIV</th>
<th>RTQR</th>
<th>TQR</th>
<th>RTQR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>0.66</td>
<td>0.66</td>
<td>0.38</td>
<td>0.16</td>
<td>0.24</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td></td>
<td>One brick</td>
<td>800</td>
<td>0.66</td>
<td>0.66</td>
<td>0.34</td>
<td>0.14</td>
<td>0.21</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.65</td>
<td>0.65</td>
<td>0.26</td>
<td>0.12</td>
<td>0.17</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>Two bricks</td>
<td>600</td>
<td>0.76</td>
<td>0.65</td>
<td>0.56</td>
<td>0.25</td>
<td>0.36</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>0.7</td>
<td>0.66</td>
<td>0.47</td>
<td>0.2</td>
<td>0.28</td>
<td>0.71</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.67</td>
<td>0.66</td>
<td>0.38</td>
<td>0.15</td>
<td>0.21</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>Cross</td>
<td>600</td>
<td>1.05</td>
<td>0.79</td>
<td>0.9</td>
<td>0.4</td>
<td>0.61</td>
<td>1.03</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.71)</td>
<td>(1.13)</td>
<td>(1.79)</td>
<td>(0.3)</td>
<td>(1.13)</td>
<td>(1.81)</td>
<td>(0.09)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>0.8</td>
<td>0.71</td>
<td>0.61</td>
<td>0.34</td>
<td>0.47</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.72</td>
<td>0.67</td>
<td>0.47</td>
<td>0.3</td>
<td>0.39</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>Pyramid</td>
<td>600</td>
<td>2.79</td>
<td>1.98</td>
<td>2.65</td>
<td>2.69</td>
<td>1.88</td>
<td>2.74</td>
<td>1.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.45)</td>
<td>(1.55)</td>
<td>(0.37)</td>
<td>(1)</td>
<td>(1.47)</td>
<td>(0.35)</td>
<td>(1.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>1.11</td>
<td>0.94</td>
<td>1</td>
<td>0.81</td>
<td>0.79</td>
<td>1.12</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.27)</td>
<td>(0.06)</td>
<td>(0.31)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.29)</td>
<td>(0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.76</td>
<td>0.65</td>
<td>0.55</td>
<td>0.45</td>
<td>0.48</td>
<td>0.75</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.35)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td></td>
<td>Cauchy</td>
<td>600</td>
<td>1.14</td>
<td>0.98</td>
<td>0.68</td>
<td>11.78</td>
<td>0.4</td>
<td>1.15</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.45)</td>
<td>(0.06)</td>
<td>(0.11)</td>
<td>(11.11)</td>
<td>(0.11)</td>
<td>(1.64)</td>
<td>(1.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>1.02</td>
<td>0.97</td>
<td>0.52</td>
<td>10.9</td>
<td>0.29</td>
<td>1.02</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.36)</td>
<td>(0.03)</td>
<td>(0.06)</td>
<td>(8.32)</td>
<td>(0.03)</td>
<td>(1.59)</td>
<td>(0.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>0.97</td>
<td>0.96</td>
<td>0.39</td>
<td>5.94</td>
<td>0.22</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(7.21)</td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>Two bricks</td>
<td>600</td>
<td>11.53</td>
<td>11.3</td>
<td>10.31</td>
<td>15.47</td>
<td>9.81</td>
<td>11.29</td>
<td>10.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.57)</td>
<td>(0.56)</td>
<td>(0.77)</td>
<td>(9.67)</td>
<td>(1.12)</td>
<td>(0.49)</td>
<td>(0.52)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>11.16</td>
<td>8.12</td>
<td>8.32</td>
<td>12.99</td>
<td>7.72</td>
<td>10.52</td>
<td>6.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.46)</td>
<td>(0.45)</td>
<td>(0.68)</td>
<td>(9.9)</td>
<td>(0.65)</td>
<td>(0.41)</td>
<td>(0.46)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>1.06</td>
<td>0.97</td>
<td>0.62</td>
<td>8.57</td>
<td>0.31</td>
<td>1.06</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.35)</td>
<td>(0.28)</td>
<td>(0.05)</td>
<td>(9.85)</td>
<td>(0.04)</td>
<td>(0.36)</td>
<td>(0.22)</td>
</tr>
<tr>
<td></td>
<td>Cauchy</td>
<td>600</td>
<td>14.24</td>
<td>13.45</td>
<td>12.92</td>
<td>15.55</td>
<td>12.6</td>
<td>14.6</td>
<td>13.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.62)</td>
<td>(1.14)</td>
<td>(0.83)</td>
<td>(11.32)</td>
<td>(0.81)</td>
<td>(1.42)</td>
<td>(1.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>11.74</td>
<td>8.84</td>
<td>8.34</td>
<td>14.82</td>
<td>6.34</td>
<td>14.29</td>
<td>8.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.42)</td>
<td>(0.77)</td>
<td>(0.77)</td>
<td>(9.27)</td>
<td>(0.72)</td>
<td>(1.36)</td>
<td>(0.73)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>1.95</td>
<td>1.17</td>
<td>0.91</td>
<td>9.44</td>
<td>0.67</td>
<td>1.63</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.67)</td>
<td>(0.31)</td>
<td>(0.69)</td>
<td>(9.08)</td>
<td>(0.4)</td>
<td>(0.57)</td>
<td>(0.33)</td>
</tr>
<tr>
<td></td>
<td>Pyramid</td>
<td>600</td>
<td>6.93</td>
<td>6.39</td>
<td>5.82</td>
<td>14.42</td>
<td>5.51</td>
<td>6.8</td>
<td>6.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.8)</td>
<td>(0.53)</td>
<td>(0.41)</td>
<td>(10.9)</td>
<td>(0.44)</td>
<td>(0.81)</td>
<td>(0.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>6.86</td>
<td>3.23</td>
<td>5.29</td>
<td>14.27</td>
<td>2.14</td>
<td>6.59</td>
<td>3.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.71)</td>
<td>(0.46)</td>
<td>(0.37)</td>
<td>(5.86)</td>
<td>(0.33)</td>
<td>(0.83)</td>
<td>(0.52)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>3.89</td>
<td>2.16</td>
<td>2.65</td>
<td>11.61</td>
<td>1.18</td>
<td>3.84</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.67)</td>
<td>(0.38)</td>
<td>(0.37)</td>
<td>(5.09)</td>
<td>(0.31)</td>
<td>(0.6)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>
matter and pial surfaces, and surface registration. See Glasser et al. (2013) for detailed descriptions.

Moreover, we crop the preprocessed images by removing the marginal regions with zero-intensity, and hence the size of the preprocessed images is $192 \times 240 \times 192$. Following Feng, Bi and Zhang (2020), we downsize each MRI to 50% of its original size, that is, $96 \times 120 \times 96$, unless stated otherwise. While leading to some loss of image resolution, this step improves computational speed and estimation efficiency, as it would considerably reduce the number of parameters to estimate given the ultrahigh dimensional images and the limited sample size. Given that our major interest is to identify active regions, we demonstrate in Section 5.2 that the performance of the proposed method is stable with varying dimensions. Our results are consistent with the practice in the literature that also suggests a relatively limited sacrifice in prediction; see, for example, Liu, Zhang and Shen (2014), Li et al. (2016) and Zhang et al. (2019).

As a final step, we standardize each voxel across the subjects. In total, we analyze 1102 participants using the tensor quantile regression framework with the primary interest of identifying brain subregions associated with quantiles of fluid intelligence. Demographic covariates of interest include age (mean: 28.8, SD: 3.7, range: 22-35), gender (45.8% male), years in education (mean: 14.9, SD: 1.8, range: 11-17), among others.

5.1. Prediction. We first evaluate the prediction performance of the proposed framework. In Figure 4(a), the histogram shows the left-skewed response of the dataset, which violates the normality assumption and even a log transformation may have limited improvement. We conduct a 5-fold cross-validation 100 times for the dataset to compare tensor quantile regression, its regularized form with the fusion penalty, as well as tensor linear regression and its regularized form SHIV. In each replication, we randomly sample 80% of the subjects as the training set, and use the other 20% of the subjects as the testing set. We fit the tensor regressions for nine ranks $r = 1, 2, \cdots, 9$, and choose the tuning parameter for the regularized methods using SIC as before.

We use median prediction of TQR or RTQR and the mean prediction of TR or SHIV, or its log transformed one whichever is better, to calculate MAEs for the testing set. The relative prediction efficiency, defined as the ratio of the MAE of TQR or RTQR and that of TR or SHIV, is reported as the evaluation criterion. Figures 4(b) and 4(c) show that the improvements of TQR or RTQR over TR or SHIV are generally more profound for higher ranks. For $r < 3$, the proposed methods are slightly better than their counterparts, but in reality the signal is hardly of such a low rank structure. When $r \geq 3$, the gains of using TQR or RTQR are remarkable. We find that the proposed methods improve their counterparts by 26% at the most in terms of prediction, see Table S.1 in the Supplementary Material (Li and Zhang, 2021). In addition, we plot MAEs of each method with $r = 3$ as an example for the testing set in Figure 4(d), demonstrating that the predictions by the proposed methods are reasonable.

Assessing the quality of predicted quantiles is not straightforward, since in reality there are no observations pertaining to the true quantiles. One commonly used evaluation criterion is the rate of the observations less than the predicted quantiles,

$$I_i(\tau) = 1(\hat{Q}_i \leq \tau), \quad \mathbb{E}\{I_i(\tau) | z_i, \lambda_i\} = \tau.$$

Here we fit RTQR with $r = 3$ as an example and $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The scatter plot of Figure 4(e) is found to be close to the identity line, indicating that the proposed method works reasonably well. Finally, we check the empirical coverage probabilities of the proposed method. Again, we show the results of RTQR with $r = 3$ in Figure 4(f).
The medians of the coverage are quite close to the theoretical ones, demonstrating the validity and generality of the proposed method. Quantifying the predicting uncertainties by providing prediction intervals makes the proposed method a potential prognostic tool since cognitive scores are commonly used biomarkers. For example, patients experience a serious decline in cognitive scores, which appear as low quantiles or fall outside of the prediction interval based on their recent MRI scans, indicating progression of cognitive impairment and a need for an early intervention.

![Graphs and images](http://example.com/graphs.png)

Fig 4: (a): Distribution of cognitive scores, the red line is the estimated density function. (b): Improvements of TQR over TR in terms of prediction. (c): Improvements of RTQR over SHIV in terms of prediction. (d): Mean absolute errors of various methods for prediction. (e): QQ-plot of the predicted quantiles and the empirical quantiles, the red line is 45 degree reference line. (f): Coverage probabilities of the prediction intervals, the red lines are the expected coverages.

5.2. Real image covariates and hypothetical responses. As in Zhou, Li and Zhu (2013) and Feng, Bi and Zhang (2020), one limitation of the real data analysis is lacking of ground truth and thus it is not clear how to evaluate the estimation performance. First, we use the true MRI images of dimension $96 \times 120 \times 96$ to fit a tensor linear regression adjusting for scalar covariates. Second, we simulate the signal with the estimated scalar parameters as well as a hypothetical tensor parameter in varying shapes. The signal locations are based on the identified regions (the prefrontal and anterior cingulate cortex) associated with human intelligence as we will show in Section 5.3. Specifically, all the synthesized signals are located in the prefrontal cortex, except for the second region of “two bricks.” The residues of the fitted tensor linear regression are added to the signal as the random noises so that we can mimic real data. These noises are heavy-tailed. We then fit various methods and compare their performance as in Section 4. Figure 5 summarizes the performance of those methods. First, VoxelQQR can only scan the shape of the brains due to its marginal fitting procedure and
hence has the largest RMSEs. Second, SHIV performs well for the “one brick” case, which has the simplest data structure, but fails to recognize the other regions of signals. Third, TQR outperforms SHIV in all cases as it can delineate the core regions of interest better and has smaller RMSE. RTQR further improves TQR, and is uniformly the best method in terms of estimation. Overall, the estimates of the proposed methods show competitive performance in a realistic setting. Furthermore, we carry out the experiment with varying dimensions and different signal locations. The MRI images are downsized to $38 \times 48 \times 38$, $58 \times 72 \times 58$, $77 \times 96 \times 77$, respectively, which correspond to approximately 20%, 30%, 40% of the original size. We find that the patterns are consistent with that of Figure 5. See Figures S.4, S.5, and S.6 in the Supplementary Material (Li and Zhang, 2021) for details. In terms of computational times, as an example, we take the “cross” shape as the true signal and repeat the learning process 10 times for various methods. The averaged run times shown in Figure 6 are recorded based on a compute node in a cluster with 8 Intel Xeon E5-2660 v3 CPU at 2.60 GHz and 64 GB of memory. The dimensions of parameters of the varying dimensions are approximately $7 \times 10^4$, $2 \times 10^5$, $6 \times 10^5$, $7 \times 10^6$. While the computational times of VoxelQR dramatically increase as the dimensions increase, other tensor-based methods are much more scalable. Note that the run times of SHIV and RTQR are based on a grid of nine points as we fix the rank and select the smoothing parameter. The results suggest that in terms of computational efficiency, RTQR is comparable to SHIV, which essentially employs a coordinate descent Lasso solver (Tibshirani and Taylor, 2011) provided by MATLAB (The MathWorks Inc., 2020), but RTQR takes less run times when the dimensions are ultrahigh; TQR is close to TR, and they take the least time since no penalty terms are imposed.

5.3. Identifying brain regions associated with quantiles of fluid intelligence. In this subsection, we identify the brain subregions associated with quantiles of fluid intelligence for the HCP. Three levels of quantiles $\tau \in \{0.25, 0.5, 0.75\}$ are considered. Following Feng, Bi and Zhang (2020), we fit the regularized tensor quantile regression with rank 7 to sufficiently capture the variation of the data, the tuning parameters are selected by SIC. The estimated coefficients are used for calculating the means and standard deviations of magnitudes so that the regions with a magnitude that is seven standard deviations larger than the mean are selected as the regions of interest (ROI). Figure 7 presents the most activated subregions of brains associated with three quantiles of fluid intelligence. Here, we use three colors to represent the regions mostly associated with different quantile levels: green is for 0.25 quantile, red is for 0.5 quantile, and blue is for 0.75 quantile. Figure 8 shows the identified brain subregions overlaid on a template. Moreover, Figure 9 and Figure S.7 in the Supplementary Material (Li and Zhang, 2021) show the magnitudes of the estimated coefficients.

We have the following findings. First, all the activated regions reside in the gray matters. This makes sense as gray matter contains most of the neuronal cell bodies and consists of brain regions related to sensory perception, memory, and decision making. Second, the HCP region of 0.25 quantile is located in frontal lobes as well as Brodmann area 10, which is highly related to general and fluid intelligence, see Haier et al. (2004); Kievit et al. (2014); Chan et al. (2018). This brain region is relevant to planning complex cognitive behavior, decision making, and moderating certain aspects of language. Orchestration of thoughts and actions in accordance with internal goals is implicated as the basic activity of this region. Third, the HCP region of 0.5 quantile is located in the insular cortex. The insula has been revealed to play a role in some higher-level functions that operate mainly in humans. Nevertheless, this region is rarely reported in human intelligence studies. A recent research conducted by Hidese et al. (2020) reveals that the volume of the gray matter region is reported as being significantly correlated with adult intelligence. Fourth, the HCP region of 0.75 quantile resides in the anterior cingulate cortex, which is in concordance with Gong et al. (2005). Higher-level functions such as anticipation of tasks, error detection, attention, and modulation of
emotional responses are attributed to this region. In summary, we identify significant subregions of the prefrontal and anterior cingulate cortex for the HCP, which are well known for their association with fluid intelligence. The discovery of the insular cortex is novel, although
Fig 6: Averaged run times (in seconds) based on 10 runs.

Fig 7: The most activated subregions associated with quantiles of human intelligence. The three columns correspond to the axial, coronal, and sagittal views of a brain template. The green region is mostly associated with 0.25 quantile; the red region is mostly associated with 0.5 quantile; and the blue region is mostly associated with 0.75 quantile.

it is also reported as a region associated with adult intelligence in a recent independent study (Hidese et al., 2020).

We verify the significance of the identified subregions by fitting linear regressions and quantile regressions using the averaged voxel intensities of each ROI as well. All the regions are significantly associated with the cognitive score. Indeed, the associated P-values of linear regressions are $8.72 \times 10^{-6}$, $9.72 \times 10^{-6}$, and $9.72 \times 10^{-6}$ for the three regions of 0.25, 0.5, and 0.75 quantiles, respectively. The P-values of quantile regressions show a similar pattern. In addition, we also try VoxelQR for the HCP. The significant voxels are so sporadic that activated regions can be hardly recognized.

6. Discussion. In this article, we propose a novel tensor quantile regression framework incorporating data structures of low-rankness and piece-wise smoothness, which is robust to distribution assumption as well as potential outliers and enhances model interpretability. The
proposed quantile regression framework translates into a sequence of regularized quantile regressions with a generalized Lasso penalty, which is computationally challenging as the objective function is not linear in the marginal vectors of parameters jointly. Furthermore, both of the check loss function and the penalty term are non-smooth, which add another layer of difficulty for the computation. Hence, we develop an efficient ADMM algorithm to estimate the parameters in an iterative manner. Extensive numerical studies are carried out to examine the numerical properties of the proposed method and its competitors. We further apply the proposed method to analyze a publicly available large-scale dataset. By taking into account extreme quantiles of cognitive scores, we successfully identify the most activated regions associated with quantiles of fluid intelligence.

The proposed method has some limitations, and it can be extended in several ways: one of the future directions is ensuring non-crossing for the estimated quantiles as in Bondell, Reich and Wang (2010), which can be achieved by enforcing non-crossing constraints in the regularized quantile regression (6) and suitably embedding it into the main loop of Algorithm...
Another potential direction is modeling multiple quantiles simultaneously and imposing smoothness constraints across quantiles as in Jiang, Wang and Bondell (2013); finally, we will further study the indirect approach as shown in (3). Our preliminary results show that its prediction performance is comparable to that of the direct approach. Nevertheless, this direction warrants future investigation.

In imaging analysis, intensity represents the shade of gray of tissue components. Intensity differences represent possible differences in the tissue component of an anatomic structure between persons, whereas similar intensity values represent similar tissue types across different persons. This kind of ideas originates from Voxel-Based Morphometry (Ashburner and Friston, 2000, 2001; Good et al., 2001), which allows for the quantification of the gray matter volumes (Winkler et al., 2010), and is commonly used to explore regional anatomical differences. Compared with the traditional voxel-wise method, our approach jointly models all the voxels simultaneously using regularized techniques to characterize tissue concentration differences for MRI scans.

Our findings have several biological implications. First, the three most activated regions are present in the gray matters, suggesting that the gray matters may contain numerous neuronal cell bodies that process and release information throughout the central nervous system. Second, the identified brain regions do not appear symmetric, suggesting that there may exist neuroanatomical differences between hemispheres. Such a phenomenon called brain asymmetry has been continuously studied through intensity-based approaches. See, for example, Watkins et al. (2001) and Luders et al. (2004). Our results provide further evidence for this phenomenon that right and left hemispheres may play different roles in higher mental processes. Third, the identified regions differ by quantiles, implying that the heterogeneity of human intelligence may be explained by different activated structures in cognition. This novel finding offers a direction for further investigation.

The piece-wise smoothness of the coefficients implies that there might be latent relationships or unknown synergies across and/or at the boundary of different brain anatomical structures.

Our model is useful as a prognostic tool for assessing the cognitive impairment progression or predicting other responses of interest using whole-brain images as biomarkers. Voxel-wise methods cannot make use of all information in a single prediction model. There are alternative methods that extract feature vectors or summary statistics from the images. These methods do not exploit the raw yet rich spatial and structural information about the brain. Without directly using the brain structural information, the results from these methods are difficult to interpret. Our model not only provides a viable tool for whole-brain based prediction, but also yields interpretable results through the identification of potential regions associated with the response of interest.

Acknowledgements. The authors thank the Editor Professor Jeffrey S. Morris, an anonymous associate editor and two reviewers for many constructive comments and helpful suggestions, which greatly improved the article. This research is supported in part by U.S. National Institutes of Health (R01HG010171 and R01MH116527). Data were provided by the Human Connectome Project (1U54MH091657) funded by the 16 NIH Institutes and Centers that Support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure.

Supplement A: Online supplement. The supplementary materials contain an alternative algorithm for solving the penalized quantile regression, additional results for simulations and real data analysis.
Supplement B: Demo code. Zip file containing code to demonstrate the proposed methods with a simulated dataset.

REFERENCES

TENSOR QUANTILE REGRESSION


