Published in final edited form as:
J Am Stat Assoc. 2017 ; 112(519): 1156–1168. doi:10.1080/01621459.2016.1194846.

Generalized Scalar-on-Image Regression Models via Total Variation

Xiao Wang and
Associate Professor of Statistics, Department of Statistics, Purdue University, West Lafayette, IN 47907

Hongtu Zhu
Professor of Biostatistics, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77230, and University of North Carolina, Chapel Hill, NC 27599

for the Alzheimer's Disease Neuroimaging Initiative

Abstract

The use of imaging markers to predict clinical outcomes can have a great impact in public health. The aim of this paper is to develop a class of generalized scalar-on-image regression models via total variation (GSIRM-TV), in the sense of generalized linear models, for scalar response and imaging predictor with the presence of scalar covariates. A key novelty of GSIRM-TV is that it is assumed that the slope function (or image) of GSIRM-TV belongs to the space of bounded total variation in order to explicitly account for the piecewise smooth nature of most imaging data. We develop an efficient penalized total variation optimization to estimate the unknown slope function and other parameters. We also establish nonasymptotic error bounds on the excess risk. These bounds are explicitly specified in terms of sample size, image size, and image smoothness. Our simulations demonstrate a superior performance of GSIRM-TV against many existing approaches. We apply GSIRM-TV to the analysis of hippocampus data obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset.

Keywords

Excess risk; Functional regression; Generalized scalar-on-image regression; Prediction; Total variation

1 Introduction

The aim of this paper is to develop generalized scalar-on-image regression models via total variation (GSIRM-TV) with scalar response and imaging and/or scalar predictors. This new development is motivated by studying the predictive value of ultra-high dimensional imaging data and/or other scalar predictors (e.g., cognitive score) for clinical outcomes

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

including diagnostic status and the response to treatment in the study of neurodegenerative and neuropsychiatric diseases, such as Alzheimer’s disease (AD)(Mu and Gage 2011). For instance, the growing public threat of AD has raised the urgency to discover and validate prognostic biomarkers that may identify subjects at greatest risk for future cognitive decline and accelerate the testing of preventive strategies. In this regard, prior studies of subjects at risk for AD have examined the utility of various individual biomarkers, such as cognitive tests, fluid markers, imaging measurements, or some individual genetic markers (e.g., APOE4 gene), to capture the heterogeneity and multifactorial complexity of AD (reviewed in Weiner et al. 2012).

Our GSIRM-TV considers the use of imaging predictor X and/or scalar predictors Z to predict scalar response Y. In practice, imaging data are often represented in the form of 2-dimensional matrix or 3-dimensional array. Assume that $X \in \mathbb{R}^{N \times N}$ is a 2-dimensional matrix of size $N \times N$ which is observed without error and $Z \in \mathbb{R}^p$ is a $p \times 1$ vector with the first component being constant one. Our GSIRM-TV assumes that Y given (X, Z) follows

$$Y|(X,Z) \sim \text{Exponential Family}(\mu,\phi) \;\text{ and }\; g(\mu)=\theta_0^T Z+\langle X,\beta_0\rangle, \quad (1)$$

where μ and ϕ are, respectively, canonical and scale parameters, $\langle U, V \rangle = \Sigma_{i,j} u_{i,j} v_{i,j}$ for $U = (u_{i,j}) \in \mathbb{R}^{N \times N}$ and $V = (v_{i,j}) \in \mathbb{R}^{N \times N}$, and $g(\cdot)$ is a known link function. Moreover, θ_0 and $\beta_0(\cdot)$ are unknown parameters of interest and $\beta_0(\cdot)$ is called the *coefficient image/function*. Throughout the paper, assume that images are observed without error. We may deal with such measurement errors in images by applying some smoothing methods to reduce error in images (Li et al. 2010).

GSIRM-TV can be regarded as an extension of the well-known functional linear model (FLM) and the high-dimensional linear model (HLM) that have been extensively studied in the literature. If we regard $\langle U, V \rangle$ as an approximation of a two-dimensional integral, then GSIRM-TV is an approximated version of FLM. The literature on FLM is too vast to summarize here. Please see the well-known monographs Ramsay and Silverman (2005) and Ferraty and Vieu (2006). The functional principal component analysis (fPCA) and various penalization methods have been developed to estimate the coefficient function. For example, the fPCA method has been discussed by James (2002), Müller and Stadtmüller (2005), Hall and Horowitz (2007), Reiss and Ogden (2007, 2010), James et al. (2009), and Goldsmith et al. (2010) and the penalized method has been studied by Crambes et al. (2009), Yuan and Cai (2010), and Du and Wang (2014). On the other hand, if we vectorize X and $\beta_0(\cdot)$ as $N^2 \times 1$ vectors, model (1) takes the form of the high dimensional generalized linear regression. To achieve sparsity in β_0, various penalization methods, such as Lasso or SCAD, have been developed. Please see Tibshirani (1996), Chen et al. (1998), Fan and Li (2001), and references therein.

Compared with FLM and HLM, a key novelty of GSIRM-TV is that the coefficient image $\beta_0(\cdot)$ in model (1) is assumed to be a piecewise smooth image with unknown jumps and edges. Such assumption not only has been widely used in the imaging literature, but also is critical for addressing various scientific questions, such as the identification of brain regions associated with AD. As an illustration, we consider a data set with $n = 300$ subjects simulated from a functional linear model which is a special case of (1). The first row of Figure 1 presents the true 64 × 64 image matrix β_0, X, and Y from the left to the right. We have vectorized X, used fPCA for FLM, Lasso for HLM, and GSIRM-TV to estimate the coefficient image and presented the estimated coefficient images in the second row of Figure 1. Unfortunately, both FLM and HLM fail to capture the main feature of the true coefficient image due to their key limitations. First, fPCA requires that β_0 be well presented by the eigenfunctions of X, whereas it is not the case according to Figure 1. Second, the existing regularization methods can have difficulty in recovering β_0, since the true coefficient image is non-sparse. Moreover, most regularization methods for FLM assume that the unknown coefficient function is one-dimensional and belongs to a smoothed function space, such as the Sobolev space, and thus they will not be able to preserve edge and boundary information for the data set presented in Figure 1. In contrast, our GSIRM-TV estimate developed in this paper can truly preserve the sharp edge of the original image.

In this paper, we make two important contributions including a new estimation method based on the total variation analysis and non-asymptotic error bounds on the risk under the framework of GSIRM-TV. The total variation analysis plays a fundamental role in various image analyses since the path-breaking works of Rudin and Osher (1994) and Rudin, Osher and Fatemi (1992). The total variation penalty has been proved to be quite efficient for preserving the boundaries and edges of images (Rudin et al. 1992). Michel et al. (2011) proposed a similar total variation method for image regression and image classification, but they focus on the development of different algorithms for the TV optimization problem. According to the best of our knowledge, this is the first paper on the development of statistical analysis of the total variation method for GSIRM-TV. The fused lasso (Tibshirani et al. 2005; Friedman et al. 2007) uses a similar penalty function. But for the 2-dimensional parameter, the fused lasso and the TV penalty can be quite different. For example, the isotropic total variation penalty uses the Euclidean norm of the first differences of the parameter, rather than the sum of the absolute values of the first differences. There are a few papers on the use of two-dimensional or three-dimensional imaging predictors in FLM (Guillas and Lai 2010; Reiss and Ogden 2010; Zhou et al. 2013; James, et al. 2009; Goldsmith et al. 2010; Gertheiss et al. 2013; Wang et al. 2014; Reiss et al. 2015), but none of them consider the piecewisely smoothed function with jumps and edges and the total variation analysis. We also derive nonasymptotic error bounds on the risk for the estimated coefficient image under the total variation penalty. We are able to obtain finite-sample bounds that are specified explicitly in terms of the sample size n, the image size $N \times N$, and the image smoothness.

The rest of the paper is organized as follows. Section 2 considers linear scalar-on-image regression model and proposes the TV optimization framework to estimate the unknown coefficient image. We also establish the nonasymptotic error bound for the prediction error. Section 3 extends linear scalar-on-image regression model to generalized scalar-on-image regression models. Section 4 examines the finite-sample performance of GSIRM-TV and compares it with several state-of-the-art methods, such as regularized matrix regression (Zhou and Li 2014). Section 5 applies GSIRM-TV to the use of the hippocampus imaging data for a binary classification problem. Future research directions are discussed in Section 6. The technical proofs of main theorems are given in the Appendix.

2 Linear scalar-on-image regression model

We start with considering a linear scalar-on-image regression model, which is the simplest case of GSIRM-TV (1), as follows:

$$Y = \langle X, \beta_0 \rangle + \varepsilon, \quad (2)$$

where ε is the random error with $\mathbb{E}(\varepsilon|X) = 0$ and $\mathbb{E}(\varepsilon^2|X) = \sigma^2$, and without loss of generality, both X and Y are assumed to be centered with $\mathbb{E}(Y) = \mathbb{E}(X) = 0$. Model (2) may be treated as a special case of FLM since discrete images are isometric to the space of piecewise-constant functions defined as

$$\mathscr{X} = \left\{ x \in L^2(\Omega) : x(u, v) = N X_{jk}, \frac{j-1}{N} \le u < \frac{j}{N}, \frac{k-1}{N} \le v < \frac{k}{N} \text{ for } 1 \le j, k \le N \right\},$$

where X_{jk} is the (j, k)–th pixel value of the image X and $\Omega = [0, 1]^2$. By treating β_0 as an integrable function in Ω, that is, $\beta_0 \in L^2(\Omega)$, model (2) can be rewritten as

$$Y = \int_0^1 \int_0^1 x(u, v) \beta_0(u, v)\, du dv + \varepsilon.$$

2.1 The space of bounded variation

Throughout the paper, it is assumed that β_0 is a function of bounded variation in Ω if the total variation of β_0 in Ω, denoted by $\|\beta_0\|_{TV}$, is finite and defined as follows:

$$\|\beta_0\|_{TV} = \sup \left\{ \int_\Omega \beta_0(u, v) \operatorname{div} f(u, v)\, du dv : f \in C_c^\infty(\Omega; R^2), |f|_\infty \le 1 \right\},$$

where $|f|_\infty = \sup_{(u,v)\in\Omega} |f(u, v)|$ and $C_c^\infty(\Omega; R^2)$ denotes the vector field with value in R^2, which is infinitely differentiable and has compact support in Ω. Moreover, $f(u, v) = (f_1(u, v), f_2(u, v))$ and div $f(u, v) = \partial_u f_1(u, v) + \partial_v f_2(u, v)$, where $\partial_u = \partial/\partial u$ and $\partial_v = \partial/\partial v$. The vector space of functions of bounded variation in Ω is denoted by BV(Ω). For example, if β_0 is differentiable in Ω, then $\|\beta_0\|_{TV}$ reduces to $\int_\Omega \sqrt{(\partial_u \beta_0)^2 + (\partial_v \beta_0)^2}\, du dv$. In this case, β_0 belongs to the Sobolev space $W^{1,1}(\mathscr{D})$, i.e., functions with integrable first order partial derivatives. However, the power of total variation in image analysis arises exactly from the relaxation of such constraints. The BV(Ω) is much larger than $W^{1,1}(\mathscr{D})$ and contains many interesting piecewise continuous functions with jumps and edges. This is exactly the advantage of using TV regularization over other familiar regularization methods used in the nonparametric literature. For example, the smoothing spline penalty term is not sensitive enough to capture sharp edges and jumps.

There are at least two additional advantages of using bounded variation functions in model (2). First, many real images with edges have small total variation since image edges usually reside in a low-dimensional subset of pixels. As an illustration, in Figure 2, the left panel displays the Shepp-Logan phantom image, while the middle and right panels show the two components of the discrete gradient of the phantom image, which have obvious sparse patterns. Second, BV(Ω) is mathematically tractable even though it contains many more functions with edges and jumps compared with $W^{1,1}(\mathscr{D})$.

2.2 Estimation

On the basis of model (2) and BV(Ω), we propose to solve the following TV minimization:

$$\begin{array}{ll}\text{minimize} & \|\beta\|_{TV} \\ \text{subject to} & \sum_{i=1}^{n}(Y_i-\langle X_i,\beta\rangle)^2 \le \lambda^2,\end{array} \quad (3)$$

where λ is a smoothing parameter, which controls the noise level. It is known that the above minimization problem is equivalent to the penalized optimization

$$\sum_{i=1}^{n}(Y_i-\langle X_i,\beta\rangle)^2+\breve{\lambda}\,\|\beta\|_{TV}, \quad (4)$$

where $\breve{\lambda}$ is a different smoothing parameter. The TV optimization has been widely used to reconstruct images in the compressive sensing literature (see e.g., Candès et al. 2006a; Candès et al. 2006b; Needell and Ward 2013). Using the TV optimization for one-dimensional regression has been studied by Mammen and van de Geer (1997) and Tibshirani (2014). Michel et al. (2011) discussed some algorithms to solve a similar optimization problem. To the best of our knowledge, nothing has been done on the statistical properties of the TV estimator for scalar-on-image regression models.

To solve the TV minimization (3) (or (4)), we treat $\beta = (\beta_{jk}) \in \mathbb{R}^{N\times N}$ as an $N\times N$ block of pixels with β_{jk} as its (j, k) element. Then, we define the discrete total variation of $\beta = (\beta_{jk}) \in \mathbb{R}^{N\times N}$. For any $\beta \in$ BV(Ω), the discrete gradient ∇ : BV(Ω) $\rightarrow \mathbb{R}^{N\times N\times 2}$ is defined by

$$(\nabla\beta)_{jk}=\begin{cases}(\beta_{j+1,k}-\beta_{jk},\beta_{j,k+1}-\beta_{jk}), & 1\le j,k\le N-1,\\ (0,\beta_{j,k+1}-\beta_{jk}), & j=N, 1\le k\le N-1,\\ (\beta_{j+1,k}-\beta_{jk},0), & 1\le j\le N-1, k=N,\\ (0,0), & k=j=N.\end{cases}$$

Based on $(\nabla\beta)_{jk} = ((\nabla\beta)_{jk,1}, (\nabla\beta)_{jk,2})^T$, the anisotropic version of the total variation norm $\|\beta\|_{TV}$ can be rewritten as

$$\|\beta\|_{TV}^{aniso} = \|\nabla\beta\|_1 = \sum_{jk}\left\{|(\nabla\beta)_{jk,1}| + |(\nabla\beta)_{jk,2}|\right\}.$$

On the other hand, its isotropic version is defined by

$$\|\beta\|_{TV}^{iso} = \sum_{jk}\|(\nabla\beta)_{jk}\|_2 = \sum_{jk}\sqrt{(\nabla\beta)_{jk,1}^2 + (\nabla\beta)_{jk,2}^2}.$$

The anisotropic and isotropic induced total variation norms are equivalent up to a factor of $\sqrt{2}$, i.e.,

$$\frac{1}{\sqrt{2}}\|\beta\|_{TV}^{iso} \le \|\beta\|_{TV}^{aniso} \le \sqrt{2}\|\beta\|_{TV}^{iso}.$$

We will write all results in terms of the anisotropic total variation seminorm, but our results also extend to the isotropic version.

Let A_X be an $n \times N^2$ design matrix such that the ith row is the vectorized X_i. With a slight abuse of notation, we use β to denote the coefficient matrix and its corresponding vector. We may rewrite (3) as the matrix form given by

$$\hat{\beta} = \arg\min \|\beta\|_{TV} \quad \text{subject to} \quad \|Y - A_X\beta\|_2 \le \lambda. \qquad (5)$$

We adapt an algorithm called TVAL3 based on the augmented Lagrangian method (Hestenes 1969; Powell 1969; Li 2013). Specifically, we solve an equivalent optimization problem given by

$$\min_{w,\beta}\sum_{l=1}^{N^2}\|w_l\|_1 \quad \text{subject to} \quad \|Y - A_X\beta\|_2 \le \lambda \ \text{ and } \ D_l\beta = w_l \text{ for all } l,$$

where D_l is an $2 \times N^2$ vector of constants associated with the discrete gradient. As an illustration, we consider a simple case with $N = 2$. In this case, we have $\beta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})^T$. We may choose

$$D_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix},$$
$$D_3 = \begin{bmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \text{ and } D_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

so that we have $D_1\beta = (\nabla\beta)_{11}$, $D_2\beta = (\nabla\beta)_{12}$, $D_3\beta = (\nabla\beta)_{21}$, and $D_4\beta = (\nabla\beta)_{22}$.

Its corresponding augmented Lagrangian function is given by

$$\mathscr{L}_A(w,\beta)=\sum_{l=1}^{N^2}\left\{\|w_l\|_1-\mathbf{v}_l^T(D_l\beta-w_l)+\frac{\alpha_l}{2}\|D_l\beta-w_l\|_2^2+\frac{\gamma}{2}\|A_X\beta-Y\|_2^2\right\},$$

where $\mathbf{v}_l$, α_l, and γ are tuning parameters. We may find the minimizer iteratively, and then the subproblem at each iteration of TVAL3 becomes $\min_{w_l,\beta}\mathscr{L}_A(w_l,\beta)$. In our algorithm, $\mathbf{v}_l$ is updated at each iteration. Moreover, α_l's and γ as smoothing parameters can be selected by using either the C_p criterion or the K-fold cross-validation (CV). However, its computational time can be long even under current computing facilities. In our numerical examples, we pre-fix the tuning parameters by setting $\alpha_l = 2^5$ for $l = 1, \ldots, N^2$ and $\gamma = 2^8$. The simplest way to choose γ is to try different values from 2^4 up to 2^{13} and compare the recovered images. The value of α_l is much less sensitive to the choice of γ. We leave tuning parameter optimization for our future research topic.

We describe the complete algorithm as follows.

Step 1. Initialize $\beta^{(0)}$ and $\mathbf{v}_l^{(0)}$;

Step 2. Given $\beta = \beta^{(k)}$ and $\mathbf{v}_l=\mathbf{v}_l^{(k)}$, we solve for $\omega_l^{(k+1)}$, $l = 1, \ldots, N^2$, by minimizing

$$\|\omega_l\|_1-\mathbf{v}_l^T(D_l\beta-\omega_l)+\frac{\alpha_l}{2}\|D_l\beta-\omega_l\|_2^2.$$

The explicit solution (component-wise) is given by

$$\omega_l=\begin{cases} D_l\beta-\frac{\mathbf{v}_l+1}{\alpha_l}, & \text{if } D_l\beta>\frac{\mathbf{v}_l+1}{\alpha_l};\\ 0, & \text{if} \frac{\mathbf{v}_l-1}{\alpha_l}\le D_l\beta\le\frac{\mathbf{v}_l+1}{\alpha_l};\\ D_l\beta-\frac{\mathbf{v}_l-1}{\alpha_l}, & \text{if } D_l\beta<\frac{\mathbf{v}_l+1}{\alpha_l}.\end{cases}$$

Step 3. Given $\omega_l=\omega_l^{(k+1)}$ and $\mathbf{v}_l=\mathbf{v}_l^{(k)}$, $l = 1, \ldots, N^2$, we solve for $\beta^{(k+1)}$ by minimizing

$$\sum_{l=1}^{N^2}\left\{-v_l^TD_l\beta+\frac{\alpha_l}{2}\|D_l\beta-\omega_l\|_2^2+\frac{\gamma}{2}\|A_X\beta-Y\|_2^2\right\}.$$

The explicit solution is given by

$$\beta^{(k+1)}=\left\{\sum_{l=1}^{N^2}(\alpha_lD_l^TD_l+\gamma A_X^TA_X)\right\}^{-1}\left\{\sum_{l=1}^{N^2}(\mathbf{v}_l^TD_l+\alpha_lD_l^T\omega_l+\gamma A_X^TY)\right\}.$$

Step 4. Given $\beta = \beta^{(k+1)}$, $\omega_l = \omega_l^{(k+1)}$, update $\mathbf{v}_l^{(k+1)}$ by using

$$\mathbf{v}_l^{(k+1)} = \mathbf{v}_l^{(k)} = \alpha_l \left(D_l \beta^{(k+1)} - \omega_l^{(k+1)} \right).$$

Step 5. Iterate Steps 2–4 until convergence.

2.3 The error bound

In this subsection, we establish the nonasymptotic error bound for the TV estimator $\hat{\beta}$ based on model (2). We consider two types of distances to measure the error. The first one is a weighted L_2 distance such that

$$\|\hat{\beta} - \beta_0\|_{X,2} = \left\{ {}^{*}(\langle X_{n+1}, \hat{\beta} - \beta_0 \rangle^2) \right\}^{1/2},$$

where $\mathbb{E}^*$ represents taking expectation with respect to (Y_{n+1}, X_{n+1}) only. The second one is the TV distance between $\hat{\beta}$ and β_0, $\|\hat{\beta} - \beta_0\|_{TV}$.

We derive both error bounds by means of Haar wavelet basis. Various wavelet bases are commonly used to effectively represent images and the Haar wavelet is the simplest possible wavelet. The bivariate Haar wavelet basis for $L_2(\Omega)$ can be constructed as follows. Let $\phi^0(t) = I_{[0,1)}$ be the indicator function, and the mother wavelet $\phi^1(t) = 1$ for $t \in [0, 1/2)$ and −1 for $t \in [1/2, 1)$. Starting from the multivariate functions

$$\phi^d(s,t) = \phi^{d_1}(s)\phi^{d_2}(t), \quad d \in \{(0,1),(1,0),(1,1)\},$$

the bivariate Haar basis functions include the indicator function $I_{[0,1)^2}$ and other functions

$$\phi_{j,k}^d(u,v) = 2^j \phi^d(2^j x - k), \;\; d \in \{(0,1),(1,0),(1,1)\}, \;\; x = (u,v), \;\; j \geq 0, k \in \mathbb{Z}^2 \cap 2^j[0,1)^2.$$

The bivariate Haar wavelet basis is an orthonormal basis for $L_2[0, 1)^2$. Note that discrete images are isometric to the space $\mathscr{I}_N \subset L_2[0, 1)^2$ of piecewise constant functions $\mathscr{I}_N = \left\{ f \in L_2[0,1)^2 : f(s,t) = c_{jk}, \frac{j-1}{N} \leq s < \frac{j}{N}, \frac{k-1}{N} \leq t < \frac{k}{N} \right\}$ via the identification $c_{jk} = NX_{jk}$. Letting $N = 2^J$, the bivariate Haar basis restricted to the N^2 basis functions $\{I_{[0,1)^2}$, $\phi_{j,k}^d$, $j \leq J-1$, $d \in \{(0, 1), (1, 0), (1, 1)\}$, $k \in \mathbb{Z}^2 \cap 2^j[0, 1)^2\}$ forms an orthonormal basis for $\mathbb{R}^{N \times N}$. Denote by Φ the discrete bivariate Haar transformation and $\{\phi_l\}$ the Haar basis, in which $\Phi\beta \in \mathbb{R}^{N \times N}$ contains the bivariate Haar wavelet coefficients of β. Next, we review a theoretical result of Petrushev et al. (1999), who proved a deep and nontrivial result on BV(Ω). Specifically, it states that the Haar wavelet coefficients of $\beta_0 \in$ BV(Ω) are in weak ℓ_1. That is, if the Haar coefficients are sorted decreasingly according to their absolute values, then the l-th rearranged coefficient is in absolute value less than $c\|\beta_0\|_{BV}/l$ with c being an absolute constant.

Invoking Haar wavelets is only for theoretical investigation and we do not estimate the Haar coefficients directly. We now introduce the main assumptions of this paper:

A1 Assume that the coefficient image β_0 in the space of $N \times N$ blocks of pixel values with bounded variation. Assume that the error ε is sub-Gaussian.

A2 Assume that the discrete Haar representation of the image predictor X is $X = \sum_l \rho_l^{1/2} \xi_l \phi_l$, where ρ_l are positive constants and ξ_l are independently and identically distributed sub-Gaussian random variables with zero mean and unit variance.

A3 For any $\beta \in \mathrm{BV}(\Omega)$, write $\beta = \Sigma_l \gamma_l \phi_l$, where the γ_l are the Haar basis coefficients of β. We arrange γ_l in a decreasing order according to their absolute values and denote the sorted coefficients as $\gamma_{(l)}$. Assume that the corresponding sorted $\rho_{(l)}$ associated with the same basis function satisfies $c_1 s^{-2q} \le \rho_{(s)} \le c_2 s^{-2q}$ with $q > 1/2$ for each s and two positive constants c_1, c_2.

Assumption A2 on the wavelet representation of X is reasonable because the discrete wavelet transformation approximately decorrelates or "whitens" data (Vidakovic 1999). Although we might use the Karhunen-Loève expansion of X, we do not adopt this approach in order to avoid additional complexity associated with the estimation of eigenfunctions. When we sort the Haar wavelet coefficients of both β and X, the corresponding basis functions may not follow the same order. Assumption A3 specifies the decay rate of the Haar wavelet coefficients of X. From A2, the predictor images X_i can be written as $X_i = \sum_l \rho_l^{1/2} \xi_{il} \phi_l$. Let $\tilde{A}$ be an $n \times N^2$ matrix with the (i, l)-th element being $\xi_{il} / \sqrt{n}$. It is well-known that $\tilde{A}$ satisfies the restricted isometry property (RIP) with a large probability (Candès et al. 2006a, 2006b). Specifically, if $n \ge C^{-2} s \log(N^2/s)$, then with probability exceeding $1 - 2e^{-Cn}$, we have

$$(1-\delta)\|u\|_2^2 \le \|\tilde{A}u\|_2^2 \le (1+\delta)\|u\|_2^2 \quad (6)$$

for all s-sparse vectors $u \in \mathbb{R}^{N^2}$ with a small RIP constant $\delta < C$.

Let $\{\hat{\gamma}_l\}$ and $\{\gamma_l\}$ be, respectively, the wavelet coefficients of $\hat{\beta}$ and β_0. It turns out that $\|\hat{\beta} - \beta_0\|_{X,2} = \{\Sigma_l \rho_l (\hat{\gamma}_l - \gamma_l)^2\}^{1/2}$, which is the weighted L_2-norm of the wavelet coefficient difference. On the other hand, since $\|\phi_l\|_{TV} \le 8$ (Needell and Ward, 2013),

$$\|\hat{\beta}-\beta_0\|_{TV} \le \sum_l |\hat{\gamma}_l - \gamma_l| \|\phi_l\|_{TV} \le 8\|\hat{\gamma}_l - \gamma_l\|_1,$$

which is bounded by the L_1-norm of the wavelet coefficient difference. We obtain the following theorem, whose detailed proof can be found in the Appendix.

Theorem 2.1—*Assumptions A1-A3 hold. Let C be an absolute constant and* $\lambda = Cn^{1/2}$. *If* $n \geq Cs^{2q+1}\log(N^2/s^{2q+1})$ *and* $\delta < 1/3$ *in* (6), *then with probability greater than* $1 - 2\exp(-Cn)$, *we have*

$$\|\hat{\beta}-\beta_0\|_{X,2} \leq C\left\{\sigma+\frac{1}{(s\log N)^{q+\frac{1}{2}}}\|\nabla\beta_0-(\nabla\beta_0)_s\|_1\right\}, \tag{7}$$

and

$$\|\hat{\beta}-\beta_0\|_{TV} \leq C\log\left(\frac{N^2}{s}\right)\left\{(s\log N)^{q+\frac{1}{2}}\sigma+\|\nabla\beta_0-(\nabla\beta_0)_s\|_1\right\}, \tag{8}$$

where $(\nabla\beta_0)_s = \arg\min_{u:s\text{-}sparse}\|\nabla\beta_0 - u\|_1$ *is the best s-sparse approximation to the discrete gradient* $\nabla\beta_0$.

Theorem 2.1 provides non-asymptotic error bounds for $\|\hat{\beta} - \beta_0\|_{X,2}$ and $\|\hat{\beta} - \beta_0\|_{TV}$, which are specified explicitly in terms of sample size n and image size $N \times N$, and the underlying smoothness of the true coefficient image based on the discrete gradient.

Remark 2.1—*We call a prediction "stable" if* $\|\hat{\beta} - \beta_0\|_{X,2} \leq C\sigma$ *holds with a high probability. Assume that the coefficient image has the sparse discrete gradient, i.e.,* $\nabla\beta_0$ *is supported on* S_0 *with* $|S_0|_0 \leq s$. *If* $\lambda = Cn^{1/2}$, *then Theorem 2.1 shows that* $\|\hat{\beta} - \beta_0\|_{X,2} \leq C\sigma$, *which indicates that our prediction procedure is stable. Furthermore, for the extreme case with noiseless data, our prediction procedure is exact. The required sample size n is of order* $s^{2q+1}\log(N^2/s^{2q+1})$, *which depends on the smoothness of the true coefficient image* β_0, *the relative smoothness between* β_0 *and X, and the image size* $N \times N$.

Remark 2.2—*The parameter q characterizes the decay rate of the wavelet coefficients of X. The larger the q, the more the required sample size. Theorem 2.1 also shows that the larger q is, the smaller the prediction error is. When* $q = 0$, *this gives the special case discussed in* Needell and Ward (2013).

3 Generalized scalar-on-image regression models

In this section, we extend all developments for model (2) to GSIRM-TV (1). Given $X \in \mathbb{R}^{N\times N}$ and $Z \in R^p$, the response variable Y is assumed to follow an exponential family distribution as

$$\exp\left(\{Y\eta(X,Z;\theta_0,\beta_0)-b(\eta(X,Z;\theta_0,\beta_0))\}/a(\psi)+c(y,\psi)\right), \tag{9}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions, and ψ is either known or considered as a nuisance parameter. Our GSIRM-TV also assumes $\beta_0 \in \mathrm{BV}(\Omega)$. It can be shown (Nelder and Wedderburn, 1972) that

$$(Y|X)=\mu(X,Z;\theta_0,\beta_0)=\dot{b}\,(\eta(X,Z;\theta_0,\beta_0))\text{ and }\mathrm{Var}(Y|X)=a(\psi)\ddot{b}(\eta(X,Z;\theta_0,\beta_0)),$$

where $\dot{b}(\eta)$ and $\ddot{b}(\eta)$ are, respectively, the first and second derivatives of $b(\eta)$ with respect to η. Moreover, $\eta(X,Z;\theta_0,\beta_0)=\dot{b}^{-1}(g^{-1}(\theta_0^T Z+\langle X,\beta_0\rangle)))$ is the canonical parameter of (9). A Gaussian distribution with variance σ^2 has $a(\psi) = \sigma^2$ and $b(\eta) = \eta^2/2$, a Bernoulli distribution has $a(\psi) = 1$ and $b(\eta) = \log(1 + e^{\eta})$, and a Poisson distribution has $a(\psi) = 1$ and $b(\eta) = e^{\eta}$.

3.1 Estimation

Let $\xi = (\theta, \beta) \in R^p \times \mathrm{BV}(\Omega)$. Given the observed data, we propose to find estimates $\hat{\xi}$ by minimizing a penalized likelihood function given by

$$n^{-1}\sum_{i=1}^{n}\{Y_i\eta(X_i,Z_i;\theta,\beta)-b(\eta(X_i,Z_i;\theta,\beta))\}+\lambda\|\beta\|_{TV}. \qquad (10)$$

We use an algorithm, which is a standard iteratively reweighted least squares for GLMs, modified to add a TV penalty, to calculate $\hat{\xi} = (\hat{\theta}, \hat{\beta})$. Given a trial estimate of ξ, denoted by $\hat{\xi}_I$, we introduce the iterative weights and the working dependent variable as

$$\hat{w}_{i,I}=\ddot{b}(\hat{\eta}_{i,I})\text{ and }\hat{Y}_{i,I}=g(\hat{\mu}_{i,I})+(Y_i-\hat{\mu}_{i,I})\,\dot{g}\,(\hat{\mu}_{i,I}), \qquad (11)$$

where $\hat{\mu}_{i,I} = \mu(X_i, Z_i; \hat{\xi}_I)$, $\dot{g}(\mu) = dg(\mu)/d\mu$, and $\hat{\eta}_{i,I} = \eta(X_i, Z_i; \hat{\xi}_I)$. Then, we can calculate the next estimate of ξ, denoted by $\hat{\xi}_{I+1}$, by minimizing

$$\hat{\xi}_{I+1}=\mathrm{argmin}_{\xi}\left\{\sum_{i=1}^{n}\omega_{i,I}[\hat{Y}_{i,I}-\partial_{\xi}\mu_{i,I}(\hat{\xi}_I)\xi]^2+\lambda\|\beta\|_{TV}\right\}, \qquad (12)$$

where $\partial_{\xi} = \partial/\partial\xi$. The optimization in (12) can be effectively solved by using TVAL3 algorithm discussed in Section 2. Finally, we can iteratively solve $\hat{\xi}_I$ until convergence.

We provide the complete algorithm as follows.

Step 1. Initialize $\xi^{(0)} = (\theta^{(0)}, \beta^{(0)})$.

Step 2. For each k, define the weights and the working dependent variable in (11), and define the objective function in (12). Use TVAL3 algorithm to solve for $\xi^{(k+1)} = (\theta^{(k+1)}, \beta^{(k+1)})$.

Step 3. Iterate Steps 2 and 3 until convergence.

We consider the logistic scalar-on-image regression model as an example. Specifically, Y_i given (X_i, Z_i) follows a Bernoulli distribution with the success probability p_i and $\text{logit}(p_i) = \langle X_i, \beta_0 \rangle + \theta_0^T Z_i$ for $i = 1, \ldots, n$. Given the current estimate $\hat{\xi}_I$, it is easy to obtain the iterative weight and effective response variable, respectively, given by

$$\hat{\omega}_{i,I} = \frac{e^{\hat{\eta}_{i,I}}}{(1+e^{\hat{\eta}_{i,I}})^2} \text{ and } \hat{Y}_{i,I} = \hat{\eta}_{i,I} + \frac{Y_i - \hat{\mu}_{i,I}}{\hat{\mu}_{i,I}(1-\hat{\mu}_{i,I})}.$$

Therefore, the estimate $\hat{\xi}_{I+1}$ can be obtained by solving a weighted penalized least squares in (12).

3.2 The error bound

We establish an non-asymptotic prediction error bound for GSIRM-TV. We need some additional assumptions as follows.

B1 Assume $\eta(X, Z; \beta_0, \theta_0)$ is bounded almost surely. Given (X, Z), the response Y is sub-Gaussian, i.e., $\mathbb{E}\{\exp(t[Y - \dot{b}(\eta(X, Z; \beta_0, \theta_0))])|X\} \leq \exp(t^2\tilde{\sigma}^2/2)$ for some $\tilde{\sigma}^2 > 0$ and all $t \in \mathbb{R}$.

B2 The function $\dot{b}(\cdot)$ is monotonic with $\inf_t \ddot{b}(t) \geq c_3$ and $\sup_t \ddot{b}(t) \leq c_4$ for two positive constants c_3 and c_4.

The sub-Gaussian assumption B1 holds for many well-known distributions, such as Gaussian. The assumption B2 requires that the second derivative of $b(\cdot)$ is bounded above and away from zero.

Theorem 3.1—*Assumptions A1–A3 and B1–B2 hold. Let* $\lambda = Cn^{-1/2}$*, where C is a positive constant. If* $n \geq Cs^{2q+1} \log (N^2/s^{2q+1})$ *and* $\delta < 1/3$ *in* (6), *with probability greater than* $1 - 2\exp(-Cn)$, *we have* $\|\hat{\theta} - \theta_0\|_2 \leq Cn^{-1/2}$,

$$\|\hat{\beta} - \beta_0\|_{X,2} \leq C\left\{1 + \frac{1}{(s\log N)^{q+\frac{1}{2}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right\}, \tag{13}$$

and

$$\|\hat{\beta}-\beta_0\|_{TV} \leq C\log\left(\frac{N^2}{s}\right)\left\{(s\log N)^{q+\frac{1}{2}}\sigma+\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\right\}. \quad (14)$$

The conditional mean of Y_{n+1} given X_{n+1} is $\dot{b}(\eta(X_{n+1}, \beta_0))$. We may measure the accuracy of $\hat{\beta}$ by $\mathbb{E}^*[\dot{b}(\eta(X_{n+1}, \hat{\beta})) - \dot{b}(\eta(X_{n+1}, \beta_0))]^2$. Under B1, this risk is bounded by $\|\hat{\beta}-\beta_0\|_{X,2}^2$ and thus, it is reasonable to study the non-asymptotic behavior of $||\hat{\beta} - \beta_0||_{X,2}$. Theorem 2.1 is a special case of Theorem 3.1 if it is assumed in Theorem 2.1 that responses follow a normal distribution. If assuming that the coefficient image has the sparse discrete gradient, Theorem 3.1 shows that $||\hat{\beta} - \beta_0||_{X,2}$ is bounded by a constant, which is proportional to σ under the assumption of Theorem 2.1. This shows that our prediction procedure is stable for GSIRM-TV.

4 Simulation Studies

In this section, we conducted a set of Monte Carlo simulations to examine the finite sample performance of the TV estimate $\hat{\beta}$ and compare it with five competing methods. The first approach (Lasso) is to calculate the Lasso estimates of β_0. The second one (Lasso-Haar) is to calculate the Lasso estimates of the Haar coefficients of β_0 and use the inverse discrete wavelet transform to calculate the estimates of β_0. The third one (Matrix-Reg) is to estimate β_0 by using a recent development called regularized matrix regression (Zhou and Li 2014), which treats the coefficient image as a matrix and penalizes the nuclear norm of this matrix. The fourth one (FPCR) is the functional principal component regression approach (Reiss and Ogden 2007, 2010) by using tensor product cubic B-splines to approximate the coefficient function. The fifth one (WNET) is to perform scalar-on-image regression in the wavelet domain by naive elastic net (Zhao et al. 2014). Among these six approaches, the TV, Lasso, Lasso-Haar, and Matrix-Reg methods have been implemented by Matlab and the FPCR and WNET methods have been implemented in the R packages ‘refund’ and ‘refund.wave’ (see Reiss et al. 2015), respectively. For the FPCR and WNET methods, we have used the default settings of both packages. The choice of wavelet basis in WNET is the Daubechies basis.

We present some results based on linear scalar-on-image regression model (2). Specifically, X_i were simulated from a 64×64 phantom map with N = 64 and 4, 096 pixels according to a spatially correlated random process $X_i = \Sigma_l l^{-q/2}\xi_{il}\phi_l$ with q = 0, 0.5, and 1, where the ξ_l are standard normal random variables and the ϕ_l are bivariate Haar wavelet basis functions. We consider four different β_0 images including triangle, oval, T-shape, and checkerboard shapes (Figure 3). Among them, the triangle and oval images are convex, while the other two are not. Errors ε_i were independently generated from $N(0, 1)$. We set n_1 = 300 for the training set and n_2 = 100 for the test set. We repeated each setting 100 times. We calculated the root mean squared prediction error (RMSPE) to compare the finite sample performance of the six different estimation methods. Let $\hat{\beta}$ be the estimated coefficient image from the training set and $\hat{Y}_i = \langle\hat{\beta}, X_i\rangle$ be the predicted responses for the test set. For each test set, RMSPE is defined by

$$\text{RMSPE}=\sqrt{n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_i-Y_i)^2}.$$

We also calculated the means and standard errors of RMSPEs for the 100 testing datasets.

Figures 4 and 5 present the estimated β_0 from a randomly selected training dataset with q = 0 and q = 0.5, respectively, for the sample size n = 300. For all four different shapes, our TV estimates can capture the sharp boundaries of the underlying shapes. In contrast, the Lasso method fails for all shapes, since the predictor images X_i are highly correlated. The Lasso estimates of the Haar coefficients can roughly capture the true shapes. However, this method cannot faithfully recover the sharp boundaries of the triangle, oval, and T shapes, whereas it does work very well for the checkerboard shape, since this checkerboard shape is exactly one of the bivariate Haar wavelet basis functions. The matrix regression approach can roughly capture the true shapes when q = 0, and unfortunately this method fails for the case when q = 0.5, for which the entries of X are spatially correlated. The PCR approach uses splines to approximate the predictor images, and it cannot preserve the sharp edges of coefficient estimator for our examples. The WNET method fails for the case when q = 0 but it can capture the shapes of the true coefficient image when the predictors are more spatially correlated.

Table 1 presents the RMSPEs of all six methods across all shapes. Overall, our TV method has significantly smaller prediction errors, in particular for q = 0. It is expected that the Lasso method leads to the largest prediction error. For all these methods, the larger q is, the smaller are their RMSPEs. For a larger q which means the predictor images are more spatially correlated, the performances of the TV, Lasso-Haar, FPCR, and WNET are similar to each other.

5 Real data analysis

To illustrate the usefulness of our proposed model, we consider anatomical MRI data collected at the baseline by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which is a large scale multi-site study collecting clinical, imaging, and laboratory data at multiple time points from healthy controls, individuals with amnestic mild cognitive impairment, and subjects with Alzheimer's disease (AD). "Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year publicprivate partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.”

Alzheimer’s disease as an age-related neurodegenerative brain disorder is often characterized by progressive loss in memory and deterioration of cognitive functions (De La Torre 2010; Weiner et al. 2012). Important neuropathological hallmarks of AD are the gradual intraneuronal accumulation of neurofibrillary tangles formed as a result of abnormal hyperphosphorylation of cytoskeletal tau protein, extracellular deposition of amyloid-β (Aβ) protein as senile plaques, and massive neuronal death. These pathologies are evident in the hippocampus, which is located in the medial temporal lobe underneath the cortical surface, and other vulnerable brain areas. The hippocampus belongs to the limbic system and plays important roles in the consolidation of information from short-term memory to long-term memory and spatial navigation (Colom et al. 2013; Fennema-Notestine et al. 2009; Luders et al. 2013).

Given the MRI scans, hippocampal substructures were segmented with FSL FIRST (Patenaude et al. 2011) and hippocampal surfaces were automatically reconstructed with the marching cube method (Lorensen and Cline 1987). We adopted a surface fluid registration based hippocampal subregional analysis package (Shi et al. 2013), which uses isothermal coodinates and fluid registration to generate one-to-one hippocampal surface registration for surface statistics computation. It introduced two cuts on a hippocampal surface to convert it into a genus zero surface with two open boundaries. The locations of the two cuts were at the front and back of the hippocampal surface. By using conformal parameterization, it essentially converts a 3D surface registration problem into a two-dimensional (2D) image registration problem. The flow induced in the parameter domain establishes high-order correspondences between 3D surfaces. Finally, various surface statistics were computed on the registered surface, such as multivariate tensor-based morphometry (mTBM) statistics (Wang et al. 2010), which retain the full tensor information of the deformation Jacobian matrix, together with the radial distance (Pizer et al. 1999). This software package and associated image processing methods have been adopted and described by various studies (Shi et al. 2014).

We applied GSIRM-TV to the hippocampus data set calculated from ADNI. The sample in our investigation includes $n = 403$ subjects: 223 healthy controls (HC) (107 females and 116 males) and 180 individuals with AD (87 females and 93 males). We consider binary disease status with 0 being HC and 1 being AD as responses. The image predictor X_i is the 2D representation of left hippocampus. The covariate vector Z_i includes constant(=1), gender (Female=0 and Male = 1), age (55–92), and behavior score (1–36). Given (X_i, Z_i), Y_i is assumed to follow a Bernoulli distribution with the success probability p_i satisfying

$$\operatorname{logit}(p_i)=\langle X_i, \beta_0\rangle+\theta_0^T Z_i \quad \text{for} \quad i=1, \ldots, n.$$

We used the iterative reweighted algorithm described above to estimate the unknown parameters.

Table 2 presents the estimates of θ_0 and their corresponding standard deviations, which were calculated by using the bootstrap method. Figure 7 shows the estimated coefficient images by using the five estimation methods. The effects around pixels (5, 40), (40, 40), (95, 40) seem to be captured well by our TV estimate. The confidence band for the coefficient image can also be obtained by using the bootstrap method. We randomly partitioned the hippocampus data set into a training set with n_1 = 203 and a test set with n_2 = 200. We repeated this random partition for 100 times and computes 100 classification errors. The average classification error of TV is 8.13% with a standard error 1.56%. We also obtain the average classification errors for other five methods. The average classification errors are 12.23%(7.36%), 21.65%(15.56%), 12.03%(11.55%), 17.13%(3.27%), 16.45%(15.57%), respectively, for Lasso, Lasso-Haar, matrix regression, FPCR, and WNET. For the WNET method, since the R code requires the image size to be a power of 2, we have added zeros to make the image size of 256 × 256 as suggested by one of the referees. Inspecting Table 2 reveals that sex and age are not significant in GSIRM-TV. We run the same procedure without sex and age and obtained a similar classification result as the full model, which is omitted from the paper.

6 Conclusion

We have developed a class of GSIRM-TVs for scalar response and imaging and/or scalar predictors, while explicitly assuming that its slope function belongs to BV(Ω). We have developed an efficient penalized total variation minimization to estimate the coefficient image. We have used simulations and real data analysis to show that GSIRM-TV is quite efficient for estimating the slope function, while preserving its edges and jumps. We have established the nonasymptotic error bound of the TV estimate for the excess risk.

It is known that many image data have small total variation and are compressible with respect to wavelet transform. Therefore, we may generalize our approach to include both total variation penalty and Lasso penalty on the wavelet coefficients. Specifically, let Φ be the wavelet transformation operator and γ be the wavelet coefficients of the coefficient image β_0. We may calculate γ by minimizing

$$\sum_{i=1}^{n}\left(Y_i-\langle X_i, \Phi^{-1}\gamma\rangle\right)^2+\lambda_1\|\Phi^{-1}\gamma\|_{TV}+\lambda_2\|\gamma\|_1, \tag{15}$$

where Φ^{-1} is the inverse discrete wavelet transform, and $\beta = \Phi^{-1}\gamma$. In (15), there are two smoothing parameters λ_1 and λ_2 which need to be selected. Efficient algorithm is also needed to be developed to solve (15). We leave this as further research work.

We have so far focused on two-dimensional (2-D) images. It would be interesting to extend our method to analyze k–dimensional (k-D) images for $k \geq 2$ (Zhou et al., 2013; Zhu et al., 2014). For example, consider a 3-D image $f \in \mathbb{R}^{N^3}$, where $f = (f_e)$, in which $e = (e_1, e_2, e_3) \in \{1, 2, 3\}^3$. The inner product can be defined as

$$\langle f, g \rangle = \sum_{e \in \{1,2,3\}^3} f_e \cdot g_e.$$

For $\ell = 1, 2$, and 3, the discrete derivative of f in the direction of r_ℓ is $f_{r_\ell} \in \mathbb{R}^{N^{\ell-1} \times N \times N^{3-\ell}}$,

$$(f_1)_e = f_{(e_1+1, e_2, e_3)} - f_{(e_1, e_2, e_3)}, (f_2)_e = f_{(e_1, e_2+1, e_3)} - f_{(e_1, e_2, e_3)}, (f_3)_e = f_{(e_1, e_2, e_3+1)} - f_{(e_1, e_2, e_3)},$$

and the 3-D discrete gradient is $(\nabla f)_e = (f_{r_\ell})_e$ for $e_\ell \leq N - 1$ and zero elsewhere. Hence the 3-D anisotropic and isotropic total variation seminorm can be defined similarly. We may consider a similar total variation optimization (4) to estimate the 3-D coefficient image. This research is currently under investigation and will be presented in another report.

Acknowledgments

The authors would like to thank Dr. YalinWang for sharing the processed data. We would also like to thank the Editor, Associate Editor, and referees for constructive comments. The research of Xiao Wang is supported by NSF grants DMS1042967 and CMMI1030246. The research of Hongtu Zhu is partially supported by NIH grants MH086633 and 1UL1TR001111, NSF grants SES-1357666 and DMS-1407655, and a grant from Cancer Prevention Research Institute of Texas. This material was based upon work partially supported by the NSF grant DMS-1127914 to the Statistical and Applied Mathematical Science Institute.

References

1. Candès W, Romberg J, Tao T. Stable signal recovery from incomplete and inaccurate measurements. Comm Pure App Math. 2006a; 59:1027–1023.
2. Candès W, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inform Theory. 2006b; 52:489–509.
3. Chen S, Donoho DL, Saunders M. Atomic decomposition for basis pursuit. SIAM J Sci Comp. 1998; 20:33–61.
4. Colom R, Stein JL, Rajagopalan P, Martńez K, Hermel D, Wang Y, Alvarez Linera J, Burgaleta M, Quiroga AM, Shih PC, Thompson PM. Hippocampal structure and human cognition: key role of spatial processing and evidence supporting the efficiency hypothesis in females. Intelligence. 2013; 41:129–140. [PubMed: 25632167]
5. Crambes C, Kneip A, Sarda P. Smoothing splines estimators for functional linear regression. Annals of Statistics. 2009; 37:35–72.
6. De La Torre JC. Alzheimers disease is incurable but preventable. Journal of Alzheimers Disease. 2010; 20:861–870.
7. Du P, Wang X. Penalized likelihood functional regression. Statistica Sinca. 2014; 24:1017–1041.
8. Efron B. How biased is the apparent error rate of a prediction rule? J Amer Statist Assoc. 1986; 81:461–470.

9. Efron B. The estimation of prediction error: Covariance penalties and cross-validation (with discussion). J Amer Statist Assoc. 2004; 99:619–642.
10. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc. 2001; 96:1348–1360.
11. Fennema-Notestine C, Hagler DJ Jr, McEvoy LK, Fleisher AS, Wu EH, Karow DS, Dale AM. Alzheimer's Disease Neuroimaging Initiative. Structural MRI biomarkers for preclinical and mild Alzheimer's disease. Human Brain Mapping. 2009; 30:3238–3253. [PubMed: 19277975]
12. Ferraty, F., Vieu, P. Nonparametric Functional Data Analysis: Theory and Practice. Springer-Verlag Inc; New York: 2006.
13. Friedman JT, Hastie H, Tibshirani R. Pathwise coordinate optimization. Annals of Applied Statistics. 2007; 1:302–332.
14. Gertheiss J, Maity A, Staicu AM. Variable selection in generalized functional linear model. Stat. 2013; 2:86–101. [PubMed: 25132690]
15. Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized functional regression. Journal of Computational and Graphical Statistics. 2010; 20:830–851.
16. Goldsmith J, Huang L, Crainiceanu CM. Smooth scalar-on-image regression via spatial Bayesian variable selection. Journal of Computational and Graphical Statistics. 2014; 23:46–64. [PubMed: 24729670]
17. Guillas S, Lai MJ. Bivariate splines for spatial functional regression models. Journal of Nonparametric Statistics. 2010; 22:477–497.
18. Hall P, Horowitz JL. Methodology and convergence rates for functional linear regression. Annals of Statistics. 2007; 35:70–91.
19. James GM. Generalized linear models with functional predictors. Journal of the Royal Statistical Society, Series B. 2002; 64:411–432.
20. James GM, Wang J, Zhu J. Functional linear regression that's interpretable. Annals of Statistics. 2009; 37:2083–2108.
21. Hestenes, MR. Multiplier and gradient methods, Journal of Optimization Theory and Applications. In: Zadeh, LA.Neustadt, LW., Balakrishnan, AV., editors. Computing Methods in Optimization Problems. Vol. 4. Academic Press; New York: 1969. p. 303-320.
22. Li Y, Wang N, Carroll RJ. Generalized functional linear models with semi parametric single-index interactions. Journal of the American Statistical Association. 2010; 105:621–633. [PubMed: 20689644]
23. Li, C. PhD Dissertation. Rice University; 2013. Compressive Sensing for 3D Data Processing Tasks: Applications, Model and Algorithms.
24. Lorensen, WE., Cline, HE. Marching cubes: a high resolution 3D surface construction algorithm. Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 87; 1987. p. 163-169.
25. Luders E, Thompson PM, Kurth F, Hong JY, Phillips OR, Wang Y, Gutman BA, Chou YY, Narr KL, Toga AW. Global and regional alterations of hippocampal anatomy in long-term meditation practitioners. Hum Brain Mapp. 2013; 34:3369–3375. [PubMed: 22815233]
26. Mammen E, van de Geer S. Locally adaptive regression splines. Annals of Statistics. 1997; 25:387–413.
27. Michel V, Gramfort A, Varoquaux G, Eger E, Thirion B. Total variation regularization for fMRI-based prediction of behavior. IEEE Transactions on Medical Imaging. 2011; 30:1328–1340. [PubMed: 21317080]
28. Mu Y, Gage F. Adult hippocampal neurogenesis and its role in Alzheimers disease. Molecular Neurodegeneration. 2011; 6:85. [PubMed: 22192775]
29. Müller HG, Stadtmüller U. Generalized functional linear models. Annals of Statistics. 2005; 33:774–805.
30. Needell D, Ward R. Stable image reconstruction using total variation minimization. SIAM J Imaging Sciences. 2013; 6:1035–1058.
31. Nelder J, Wedderburn R. Generalized linear models. Journal of the Royal Statistical Society, Series A. 1972; 135:370–384.

32. Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage. 2011; 56:907–922. [PubMed: 21352927]

33. Petrushev PP, Cohen A, Xu H, DeVore R. Nonlinear approximation and the space BV(R^2). American Journal of Mathematics. 1999; 121:587–628.

34. Powell, MJD. A Method for Nonlinear Constraints in Minimization Problems. In: Fletcher, R., editor. Optimization. Academic Press; London, New York: 1969. p. 283-298.

35. Ramsay, JO., Silverman, BW. Functional Data Analysis. Springer-Verlag Inc; New York: 2005.

36. Reiss PT, Huo L, Zhao Y, Kelly C, Ogden RT. Wavelet-domain regression and predictive inference in psychiatric neuroimaging. Annals of Applied Statistics. 2015; 9(2):1076–1101. [PubMed: 27330652]

37. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. Journal of the American Statistical Association. 2007; 102:984–996.

38. Reiss PT, Ogden RT. Functional generalized linear models with images as predictors. Biometrics. 2010; 66:61–69. [PubMed: 19432766]

39. Rudin LI, Osher S. Total variation based image restoration with free local constraints. Proc 1st IEEE ICIP. 1994; 1:31–35.

40. Rudin LI, Osher S, Fatemi E. Nonlinear total variation noise removal algorithm. Physica D. 1992; 60:259–268.

41. Shi J, Lepore N, Gutman B, Thompson PM, Baxter L, Caselli RJ, Wang Y. Genetic influence of APOE4 genotype on hippocampal morphometry - an N=725 surface-based ADNI Study. Hum Brain Mapp. 2014; 35:3902–3918.

42. Shi J, Thompson PM, Gutman B, Wang Y. Surface fluid registration of conformal representation: application to detect disease burden and genetic influence on hippocampus. NeuroImage. 2013; 78:111–134. [PubMed: 23587689]

43. Tibshirani R. Regression shrinkage and selection via the Lasso. J of Royal Statis Soc B. 1996; 58:267–288.

44. Tibshirani R. Adaptive piecewise polynomial estimation via trend filtering. Annals of Statistics. 2014; 42:285–323.

45. Tibshirani R, Taylor J. Degrees of freedom in Lasso problems. Annals of Statistics. 2012; 40:1198–1232.

46. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. JR Statist SocB. 2005; 67:91–108.

47. van der Vaart, AW., Wellner, JA. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer; New York: 1996.

48. Vidakovic, B. Statistical Modeling by Wavelets. Wiley; New York: 1999.

49. Wang X, Nan B, Zhu J, Koppe R. ADNI. Regularized 3D functional regression for brain image data via Haar wavelets. Annals of Applied Statistics. 2014; 8:1045–1064. [PubMed: 26082826]

50. Wang Y, Zhang J, Gutman B, Chan TF, Becker JT, Aizenstein HJ, Lopez OL, Tamburo RJ, Toga AW, Thompson PM. Multivariate tensor based morphometry on surfaces: Application to mapping ventricular abnormalities in HIV/AIDS. NeuroImage. 2010; 49:2141–2157. [PubMed: 19900560]

51. Weiner MW, Veitcha DP, Aisen PS, Beckett LA, Cairnsh NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykino AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ. ADNI. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. Alzheimers Dement. 2012; 8:S1–S68. [PubMed: 22047634]

52. Yuan M, Cai TT. A reproducing kernel Hilbert space approach to functional linear regression. Annals of Statistics. 2010; 38:3412–3444.

53. Zhao Y, Ogden RT, Reiss PT. Wavelet-based LASSO in functional linear regression. Journal of Computational and Graphical Statistics. 2014; 21:600–617.

54. Zhou H, Li L. Regularized matrix regression. Journal of Royal Statistical Society Series B. 2014; 76:463–483.

55. Zhou H, Li L, Zhu H. Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association. 2013; 108:540–552. [PubMed: 24791032]

56. Zhu H, Fan J, Kong L. Spatially varying coefficient model for neuroimaging data with jump discontinuities. Journal of the American Statistical Association. 2014; 109:977–990.

7 Appendix

In this appendix, we provide proofs of Theorems 2.1 and 3.1. The constant C represents a universal constant independent of everything else, but it may be different from different lines.

7.1 Proof of Theorem 2.1

We prove the theorem by extending the arguments from Candès, Romberg, and Tao (2006a, 2006b) and Needell and Ward (2013). In the following, $a \lesssim b$ means that there exists a constant C such that $a \leq Cb$.

Recall that $\{\phi_l\}$ is a set of discrete bivariate Haar wavelet basis functions. Write $X_i=\sum_l \xi_{il}\rho_l^{1/2}\phi_l$, $\beta_0 = \Sigma_l\, \gamma_l\phi_l$, and $\hat{\beta} = \Sigma_l\, \hat{\gamma}_l\phi_l$. We aim to derive the error bounds of $\Sigma_l\, \rho_l(\gamma_l - \hat{\gamma}_l)^2$ and $\Sigma_l|\gamma_l - \hat{\gamma}_l|$. Denote by $\alpha = \beta_0 - \hat{\beta}$ and write $\alpha = \Sigma_l\, h_l\phi_l$, where the $h_l = \gamma_l - \hat{\gamma}_l$ are the wavelet coefficients of the difference between the true coefficient image and the estimated coefficient image. We may sort h_l in descending order according to their absolute values. Denote the sorted coefficients by $h_{(l)}$. The corresponding ρ_l with the same basis function with h_l is denoted by $\rho_{(l)}$. Note that the $\rho_{(l)}$ are not necessarily sorted, but it is assumed to satisfy Condition A3.

Let S denote the support of s largest entries in the absolute values of α. As shown in Lemma 9 of Needell and Ward (2013), the set K of wavelets which are non-constant over S has cardinality at most $8s \log N$. With an abuse of notation, let $K = 8s \log N$. Lemma 7.1 derives cone constraints for the wavelet coefficients $h_{(j)}$ and the weighted wavelet coefficients $\rho_{(j)}^{1/2}h_{(j)}$.

In the following, we focus on $\rho_{(l)}^{1/2}h_{(l)}$ for $l = K + 1, \ldots, N^2$. Let

$$\tilde{s}=cs^{2q+1}(\log N)^{2q+1}, \quad d=\left\lfloor N^2/(4\tilde{s})\right\rfloor.$$

We may write $K^c = K_1 \cup K_2 \cup \cdots \cup K_d$, where K_1 consists of $4\tilde{s}$ largest $|h_{(l)}|$ within K^c, K_2 consists of next $4\tilde{s}$ largest-magnitude of $|h_{(l)}|$, and so on. Since $\rho_{(l)}$ is of order l^{-2q} and the magnitude of each $\rho_l^{1/2}|h_l|$ in K_{j-1} is larger than that in K_j up to a constant, we have

$$\left(\sum_{l\in K_j}\rho_l|h_l|^2\right)^{1/2} \lesssim \frac{1}{2\sqrt{\tilde{s}}}\sum_{l\in K_{j-1}}\rho_l^{1/2}|h_l| \;\; \text{for} \;\; j=2,3,\ldots.$$

Combining this result with Lemma 7.1 yields

$$
\begin{aligned}
\sum_{j=2}^{d}\left(\sum_{l\in K_j}\rho_l|h_l|^2\right)^{1/2} &\lesssim \frac{1}{2\sqrt{\tilde{s}}}\sum_{l=K+1}^{N^2}\rho_{(l)}^{1/2}|h_{(l)}| \\
\lesssim \frac{1}{2\sqrt{\tilde{s}}}&\left\{s^q(\log N)^q\sum_{l=1}^{K}\rho_{(l)}^{1/2}|h_{(l)}|+\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\right\} \\
&\lesssim \frac{1}{2\sqrt{K}}\sum_{l=1}^{K}\rho_{(l)}^{1/2}|h_{(l)}|+\frac{1}{2\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1 \\
&\le \frac{1}{2}\left(\sum_{j=1}^{K}\rho_{(l)}|h_{(l)}|^2\right)^{1/2}+\frac{1}{2\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1. \qquad (16)
\end{aligned}
$$

Recall that $\hat{\beta}$ is calculated by solving (5). Let $\tilde{A}$ be an $n \times N^2$ matrix with the (i, l)th element being $n^{-1/2}\xi_{il}$, ρ be a diagonal matrix with the l-th diagonal element ρ_l, and γ and h be the wavelet coefficients of β_0 and α, respectively. Therefore, $A_X\beta_0=\sqrt{n}\tilde{A}\rho^{1/2}\gamma$ and $A_X\delta\beta=\sqrt{n}\tilde{A}\rho^{1/2}h$. Let $\lambda=C\sqrt{n}\sigma$. With probability more than $1 - e^{-Cn}$, $\|Y-A_X\beta_0\|_2 \le C\sqrt{n}\sigma$. This gives

$$
\sqrt{n}\|\tilde{A}\rho^{1/2}h\|_2=\|A_X\,\alpha\|_2=\|A_X\beta_0-A_X\hat{\beta}\|_2 \le \|Y-A_X\beta_0\|_2+\|Y-A_X\hat{\beta}\|_2 \lesssim \sqrt{n}\sigma. \qquad (17)
$$

Following the argument in Candès et al. (2006a, 2006b), if $n \ge C^{-2}s\log(N^2/s)$, then $\tilde{A}$ satisfies the restricted isometry property (RIP) with a large probability: with probability exceeding $1 - 2e^{-C\delta^2 n}$,

$$
(1-\delta)\|u\|_2^2 \le \|\tilde{A}u\|_2^2 \le (1+\delta)\|u\|_2^2, \qquad (18)
$$

for all s-sparse vector $u \in \mathbb{R}^{N^2}$. Therefore,

$$
\begin{aligned}
\sqrt{n}\sigma \rhd \sqrt{n}\|\tilde{A}\rho^{1/2}h\|_2 &\ge \sqrt{n}\|\tilde{A}(\rho^{1/2}h)_K+\tilde{A}(\rho^{1/2}h)_{K_1}\|_2-\sqrt{n}\sum_{j=2}^{d}\|\tilde{A}(\rho^{1/2}h)_{K_j}\|_2 \\
&\ge \sqrt{n(1-\delta)}\|(\rho^{1/2}h)_K+(\rho^{1/2}h)_{K_1}\|_2-\sqrt{n(1+\delta)}\sum_{j=2}^{d}\|(\rho^{1/2}h)_{K_j}\|_2 \\
&\ge \sqrt{n(1-\delta)}\|(\rho^{1/2}h)_K+(\rho^{1/2}h)_{K_1}\|_2-\sqrt{n(1+\delta)}\left(\frac{1}{2}\|(\rho^{1/2}h)_K\|_2+\frac{1}{2\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\right) \\
&\ge \left(\sqrt{1-\delta}-\frac{\sqrt{1+\delta}}{2}\right)\sqrt{n}\|(\rho^{1/2}h)_K+(\rho^{1/2}h)_{K_1}\|_2-\frac{\sqrt{n(1+\delta)}}{2\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1.
\end{aligned}
$$

Since $\delta < 1/3$, we have

$$
\sqrt{n}\|(\rho^{1/2}h)_K+(\rho^{1/2}h)_{K_1}\|_2 \lesssim 5\sqrt{n}\sigma+\frac{3\sqrt{n}}{\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1. \qquad (19)
$$

Further, it follows from (16) and (19) that

$$\|\sum_{j=2}^{d}(\rho^{1/2}h)_{K_j}\|_2 \leq \sum_{j=2}^{d}\|(\rho^{1/2}h)_{K_j}\|_2 \leq \frac{1}{2}\|(\rho^{1/2}h)_K+(\rho^{1/2}h)_{K_1}\|_2+\frac{\sqrt{n}}{2\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1.$$

We arrive at

$$\sqrt{n}\|\rho^{1/2}h\|_2 \lesssim 8\sqrt{n}\sigma+\frac{5\sqrt{n}}{\sqrt{\tilde{s}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1 \lesssim \sqrt{n}\sigma+\frac{\sqrt{n}}{(s\log N)^{q+\frac{1}{2}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1.$$

This gives

$$\|\rho^{1/2}h\|_2 \leq C\left\{\sigma+\frac{1}{(s\log N)^{q+\frac{1}{2}}}\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\right\},$$

which provides the weighted L_2 error bound.

Finally, because

$$\begin{aligned}\sqrt{n}\sigma \rhd \sqrt{n}\|\tilde{A}\rho^{1/2}h\|_2 &\geq \sqrt{n(1-\delta)}\|(\rho^{1/2}h)\|_2\\ \geq \sqrt{n(1-\delta)}\|(\rho^{1/2}h)_K\|_2 &\geq \sqrt{n(1-\delta)}K^{-q}\Big(\sum_{j=1}^{K}|h_{(j)}|^2\Big)^{1/2},\end{aligned}$$

we have

$$\left(\sum_{j=1}^{K}|h_{(j)}|^2\right)^{1/2} \lesssim (s\log N)^q\sigma.$$

Combining this with (20) leads to the L_1 error bound since

$$\begin{aligned}\sum_{j=1}^{N^2}|h_{(j)}| &\leq (1+\log(\tfrac{N^2}{s}))\sum_{j=1}^{K}|h_{(j)}|+\log\left(\tfrac{N^2}{s}\right)\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\\ &\leq (1+\log(\tfrac{N^2}{s}))K^{1/2}\Big(\sum_{j=1}^{K}|h_{(j)}|^2\Big)^{1/2}+\log\left(\tfrac{N^2}{s}\right)\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\\ &\lesssim (1+\log(\tfrac{N^2}{s}))K^{q+\frac{1}{2}}\sigma+\log\left(\tfrac{N^2}{s}\right)\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\\ &\lesssim \log\left(\tfrac{N^2}{s}\right)\left\{(s\log N)^{q+\frac{1}{2}}\sigma+\|\nabla\beta_0-(\nabla\beta_0)_S\|_1\right\}.\end{aligned}$$

This completes the proof of the theorem.

Lemma 7.1

Let $K = 8s \log N$. Then

$$\sum_{j=K+1}^{N^2} |h_{(j)}| \lesssim \log\left(\frac{N^2}{s}\right)\left(\sum_{j=1}^{K} |h_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_S\|_1\right), \tag{20}$$

$$\sum_{j=K+1}^{N^2} \rho_{(j)}^{1/2} |h_{(j)}| \lesssim s^q (\log N)^q \sum_{j=1}^{K} \rho_{(j)}^{1/2} |h_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_S\|_1. \tag{21}$$

Proof—Let $\alpha = \beta_0 - \hat{\beta}$. We first derive a cone constraint for α. Let $\tilde{S}$ be the support of the largest s elements of $\nabla\beta_0$. Observe that

$$\|\nabla\hat{\beta}\|_1 \le \|\nabla\beta_0\|_1 = \|(\nabla\beta_0)_{\tilde{S}}\|_1 + \|(\nabla\beta_0)_{\tilde{S}^c}\|_1,$$

and on the other hand,

$$\begin{aligned}\|\nabla\hat{\beta}\|_1 &= \|(\nabla\beta_0)_{\tilde{S}} - (\nabla\alpha)_{\tilde{S}}\|_1 + \|(\nabla\beta_0)_{\tilde{S}^c} - (\nabla\alpha)_{\tilde{S}^c}\|_1 \\ &\ge \|(\nabla\beta_0)_{\tilde{S}}\|_1 - \|(\nabla\alpha)_{\tilde{S}}\|_1 - \|(\nabla\beta_0)_{\tilde{S}^c}\|_1 + \|(\nabla\alpha)_{\tilde{S}^c}\|_1.\end{aligned}$$

Combining these two inequalities yields

$$\|(\nabla\alpha)_{\tilde{S}^c}\|_1 \le \|(\nabla\alpha)_{\tilde{S}}\|_1 + 2\|\nabla\beta_0 - (\nabla\beta_0)_{\tilde{S}}\|_1. \tag{22}$$

The cone constraint on the discrete gradient can be transferred to a cone constraint on the wavelet coefficients. Write

$$\alpha = \sum_{j \in S} h_j \phi_j + \sum_{j \in S^c} h_j \phi_j,$$

where the wavelet coefficients are nonconstant over S which has cardinality at most $K = 8s \log N$. Recall that $|h_j| \le Cj^{-1}\|\nabla\alpha\|_1$. From (22) we have

$$
\begin{aligned}
\sum_{j=K+1}^{N^2}|h_{(j)}| \le \sum_{j=s+1}^{N^2}|h_{(j)}| &\lesssim \log\left(\frac{N^2}{s}\right)\|\nabla\alpha\|_1\\
&=\log\left(\frac{N^2}{s}\right)\left(\|(\nabla\alpha)_{\tilde{S}}\|_1+\|(\nabla\alpha)_{\tilde{S}^c}\|_1\right)\\
&\lesssim \log\left(\frac{N^2}{s}\right)\left(2\|(\nabla\alpha)_{\tilde{S}}\|_1+2\|\nabla\beta_0-(\nabla\beta_0)_{\tilde{S}}\|_1\right)\\
&\lesssim \log\left(\frac{N^2}{s}\right)\left(\sum_{j=1}^{K}|h_{(j)}|+\|\nabla\beta_0-(\nabla\beta_0)_{S}\|_1\right),
\end{aligned}
$$

where the last inequality holds because $\|\nabla\phi_j\|_1 \le 8$ (Needell and Ward, 2013), and

$$
\|(\nabla\alpha)_{\tilde{S}}\|_1=\|\nabla(\sum_{j\in\tilde{S}}h_{(j)}\phi_{(j)})\|_1 \le \sum_{j\in\tilde{S}}|h_{(j)}|\|\nabla\phi_{(j)}\|_1 \le 8\sum_{j\in\tilde{S}}|h_{(j)}| \le 8\sum_{j=1}^{K}|h_{(j)}|.
$$

Furthermore, since $\rho_{(j)}$ is of order j^{-2q} for $q > 0$ from A3, we have

$$
\begin{aligned}
\sum_{j=K+1}^{N^2}\rho_{(j)}^{1/2}|h_{(j)}| \lesssim \sum_{j=K+1}^{N^2}j^{-q}j^{-1}\|\nabla\alpha\|_1 &\le \|\nabla\alpha\|_1\\
&\lesssim \sum_{j=1}^{K}|h_{(j)}|+\|\nabla\beta_0-(\nabla\beta_0)_{S}\|_1\\
&\lesssim K^q\sum_{j=1}^{K}\rho_{(j)}^{1/2}|h_{(j)}|+\|\nabla\beta_0-(\nabla\beta_0)_{S}\|_1\\
&=s^q(\log N)^q\sum_{j=1}^{K}\rho_{(j)}^{1/2}|h_{(j)}|+\|\nabla\beta_0-(\nabla\beta_0)_{S}\|_1,
\end{aligned}
\tag{23}
$$

which completes the proof of the lemma.

7.2 Proof of Theorem 3.1

Write $\xi = (\theta, \beta)$, $W = (X, Z)$, and $W_i = (X_i, Z_i)$, $i = 1, \ldots, n$. Denote $\eta_i(\xi) = \eta(W_i; \xi)$, $\eta_W(\xi) = \eta(W; \xi)$, and

$$
M_n(\xi)=-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta_i(\xi)-b(\eta_i(\xi))\}, \quad \overline{M}_n(\xi)=-\frac{1}{n}\sum_{i=1}^{n}\left\{\dot{b}\,(\eta_i(\xi))\eta_i(\xi)-b(\eta_i(\xi))\right\}.
$$

Recall that $\hat{\xi}$ minimizes $-M_n(\xi) + \lambda\|\beta\|_{TV}$. We have

$$
0 \le M_n(\hat{\xi})-M_n(\xi_0)-\lambda\|\hat{\beta}\|_{TV}+\lambda\|\beta_0\|_{TV}.
$$

Direct calculation yields

$$\begin{aligned}M_n(\xi)-M_n(\xi_0)=&-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\dot{b}\;(\eta_i(\xi_0))\right)\eta_i(\xi-\xi_0)+\dot{b}\;(\eta_i(\xi_0))\eta_i(\xi-\xi_0)-(b(\eta_i(\xi))-b(\eta_i(\xi_0)))\\ =&H_n(\hat{\xi})-\frac{1}{2n}\sum_{i=1}^{n}\ddot{b}(\eta_i^*)\eta_i^2(\xi-\xi_0).\end{aligned} \tag{24}$$

where $\eta_i(\xi-\xi_0)=(\theta-\theta_0)^T Z+\langle X,\beta-\beta_0\rangle$, and

$$H_n(\xi)=-\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\dot{b}\;(\eta_i(\xi_0))\right)\eta_i(\xi-\xi_0), \tag{25}$$

and η_i^* is a number between $\eta_i(\xi)$ and $\eta_i(\xi_0)$. Therefore,

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^{n}\ddot{b}(\eta_i^*)\eta_i^2(\xi-\xi_0)&\leq H_n(\hat{\xi})-\lambda\|\hat{\beta}\|_{TV}+\lambda\|\beta_0\|_{TV}\\ &\leq\sup_{\xi}|H_n(\xi)|-\lambda\|\hat{\beta}\|_{TV}+\lambda\|\beta_0\|_{TV}\\ &\lesssim n^{-1/2}+\lambda,\end{aligned}$$

where the last inequality is because of Lemma 7.2.

Let $g^*=(g_1^*(X_1,\ldots,X_n),\ldots,g_n^*(X_1,\ldots,X_n))^T$ be the least favorable direction such that, for any $g=(g_1(X_1,\ldots,X_n),\ldots,g_n(X_1,\ldots,X_n))^T$, we have

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i-g_i^*(X_1,\ldots,X_n))^T g_i(X_1,\ldots,X_n)=0.$$

Note that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\eta_i^2(\hat{\xi}-\xi_0)=&\frac{1}{n}\sum_{i=1}^{n}\left((\hat{\theta}-\theta_0)^T Z_i+\langle X_i,\hat{\beta}-\beta_0\rangle\right)^2\\ =&\frac{1}{n}\sum_{i=1}^{n}\left((\hat{\theta}-\theta_0)^T(Z_i-g_i^*)+(\hat{\theta}-\theta_0)^T g_i^*+\langle X_i,\hat{\beta}-\beta_0\rangle\right)^2\\ =&(\hat{\theta}-\theta_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i-g_i^*)(Z_i-g_i^*)^T\right)(\hat{\theta}-\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\left((\hat{\theta}-\theta_0)^T g_i^*+\langle X_i,\hat{\beta}-\beta_0\rangle\right)^2.\end{aligned}$$

Assuming that $n^{-1}\sum_{i=1}^{n}(Z_i-g_i^*)(Z_i-g_i^*)^T$ is non-singular, we conclude that $\|\hat{\theta}-\theta_0\|_2\lesssim n^{-1/2}$ by choosing $\lambda=Cn^{-1/2}$, and $\|A_X\hat{\beta}-A_X\beta_0\|_2\lesssim\sqrt{n}$, which gives an equation similar to (17). The following proofs will go through with the same arguments as for the proof of Theorem 2.1. This completes the proof of Theorem 3.1.

Lemma 7.2

Assume B1–B2 hold. Let $r = q - 1/2 > 0$. *Let* $B_\delta = \{\xi: \delta/2 \le d_W(\xi, \xi_0) \le \delta\}$, *where*

$$d_W^2(\xi,\xi_0)=\|\theta-\theta_0\|_2^2+\mathbb{E}_X\left(\langle X,\beta-\beta_0\rangle^2\right)$$

Then,

$$\sup_{\xi\in B_\delta}|H_n(\xi)|=O_p\left(n^{-\frac{1}{2}}\delta^{\frac{2r+1}{2(r+1)}}\right). \tag{26}$$

Proof—Recall that $H_n(\xi) = -(P_n - \mathbb{P}) f_\xi(W, Y)$, where

$$f_\xi(W,Y)=\left(Y-\dot{b}\left(\eta_W(\xi_0)\right)\right)\eta_W(\xi-\xi_0).$$

Consider $\mathscr{M}_\delta = \{f_\xi(W, Y) : \xi \in B_\delta\}$, with $L_2(P)$ norm, i.e., for any $f \in \mathscr{M}_\delta$, $\|f\|_{P,2}^2=\mathbb{E}_{W,Y}f^2(W,Y)$.

Let $\mathbb{G}_n=\sqrt{n}(P_n-\mathbb{P})$ and $\|G_n\|_{\mathscr{M}_\delta} = \sup_{f\in\mathscr{M}_\delta}|\mathbb{G}_n f|$. Then,

$$\sup_{\xi\in B_\delta}|H_n(\xi)|=n^{-1/2}\|G_n\|_{\mathscr{M}_\delta}.$$

Therefore, it suffices to show that

$$\left\|\|\mathbb{G}_n\|_{\mathscr{M}_\delta}\right\|_{P,2}=O(\delta^{\frac{2r+1}{2(r+1)}}).$$

We prove this result by using Theorem 2.14.1 of van der Vaart and Wellner (1996) and exploiting the covering numbers of $\mathscr{M}_\delta$. For statistical applications of covering numbers, please see Chapter 2 of van der Vaart and Wellner (1996). This result can be achieved by showing that

$$\log\mathscr{N}\left(\varepsilon,\mathscr{M}_\delta,\|\cdot\|_{P,2}\right)\lesssim\varepsilon^{-1/r}\log\left(\frac{4\delta+\varepsilon}{\varepsilon}\right).$$

Suppose that there exist $\xi_1, \ldots, \xi_m \in B_\delta$ such that, for any $\xi \in B_\delta$, $\min_{1 \le i \le m} \|\xi - \xi_i\|_{P,2} \le \varepsilon$. Observe that

$$\min_{1\leq i\leq m} \mathbb{E}_{W,Y}\left[\left(Y-\dot{b}\left(\eta_W\left(\xi_0\right)\right)\right)\eta_W\left(\xi-\xi_0\right)-\left(Y-\dot{b}\left(\eta_W\left(\xi_0\right)\right)\right)\eta_W\left(\xi_i-\xi_0\right)\right]^2 \leq C\varepsilon^2.$$

Therefore, the cover number for $\mathscr{M}_\delta$ is of the same order of that for B_δ, and specifically,

$$\mathscr{N}\left(\varepsilon, \mathscr{M}_\delta, \|\cdot\|_{P,2}\right)=\mathscr{N}\left(\frac{\varepsilon}{C}, B_\delta, d_W\right).$$

Write $\beta = \Sigma_k \gamma_k \phi_k$, $\beta_0=\sum_k \gamma_k^0 \phi_k$, and $X=\sum_k \rho_k^{1/2} \xi_k \phi_k$. Hence, $d_W^2\left(\xi, \xi_0\right)=\|\theta-\theta_0\|_2^2+\sum_k \rho_k\left(\gamma_k-\gamma_k^0\right)^2$. For any $\beta = \Sigma_k \gamma_k \phi_k$, $\in B_\delta$, let $\beta^* = \Sigma_{k \le M} \gamma_{(k)} \phi_{(k)}$ with $M = \varepsilon^{-1/(r+1)}$. Since $\rho_{(l)}$ is of order l^{-2q}, $\Sigma_{\ell > s} \rho_{(\ell)}$ is of order s^{-2r} with $r = q - 1/2 > 0$. Therefore,

$$\left(\langle X, \beta-\beta^*\rangle^2\right)=\sum_{k>M} \gamma_{(k)}^2 \rho_{(k)} \lesssim \frac{1}{M^2} M^{-2r}=M^{-2(r+1)}=\varepsilon^2.$$

So if we can find $\beta_1^*, \ldots, \beta_m^* \in B_\delta^*$, where $B_\delta^*=\left\{\beta=\sum_{k\leq M} \gamma_{(k)} \phi_{(k)}: \sum_{k\leq M} \rho_{(k)}\left(\gamma_{(k)}-\gamma_{(k)}^0\right)^2 \leq \delta^2\right\}$, satisfying for all $\beta^* \in B_\delta^*$,

$$\min_{1\leq k\leq m}\left(\langle X, \beta^*-\beta_k^*\rangle^2\right) \leq \varepsilon^2,$$

it also guarantees that, for any $\beta \in B_\delta$,

$$\min_{1\leq k\leq m}\left(\langle X, \beta-\beta_k^*\rangle^2\right) \lesssim \min_{1\leq k\leq m}\left\{\left(\langle X, \beta-\beta^*\rangle^2\right)+\left(\langle X, \beta^*-\beta_k^*\rangle^2\right)\right\} \lesssim \varepsilon^2.$$

In addition, since $\theta \in \Theta$ which is a bounded subset of $\mathbb{R}^p$, we may find $\theta_1, \ldots, \theta_{\tilde{m}}$ such that, for any $\theta \in \Theta$, $\min_{1\leq k\leq \tilde{m}}\|\theta-\theta_k\|_2^2 \leq \varepsilon^2$.

Since it is known that the covering number for a ball in $\mathbb{R}^M$ is $\mathscr{N}\left(\varepsilon, B_\delta^*, d_X\right) \lesssim \{(4\delta+\varepsilon)/\varepsilon\}^M$, it follows from the above arguments that

$$\log \mathscr{N}\left(\varepsilon, \mathscr{M}_\delta, \|\cdot\|_{P,2}\right) \lesssim (M+p)\log\left(\frac{4\delta+\varepsilon}{\varepsilon}\right) \lesssim \varepsilon^{-\frac{1}{r+1}}\log\left(\frac{4\delta+\varepsilon}{\varepsilon}\right).$$

We calculate $J(1, \mathscr{M}_\delta)$ by

$$J(1,\mathscr{M}_\delta)=\int_0^1\sqrt{1+\log\mathscr{N}\left(\varepsilon,\mathscr{M}_\delta,\|\cdot\|_{P,2}\right)}d\varepsilon\lesssim\delta^{\frac{2r+1}{2(r+1)}}.$$

It follows from Theorem 2.14.1 in van der Vaart and Wellner (1996), we have $\|\|\mathbb{G}_n\|_{\mathscr{M}_\delta}\|_{P,2}=O(\delta^{\frac{2r+1}{2(r+1)}})$, which completes the proof of the lemma.

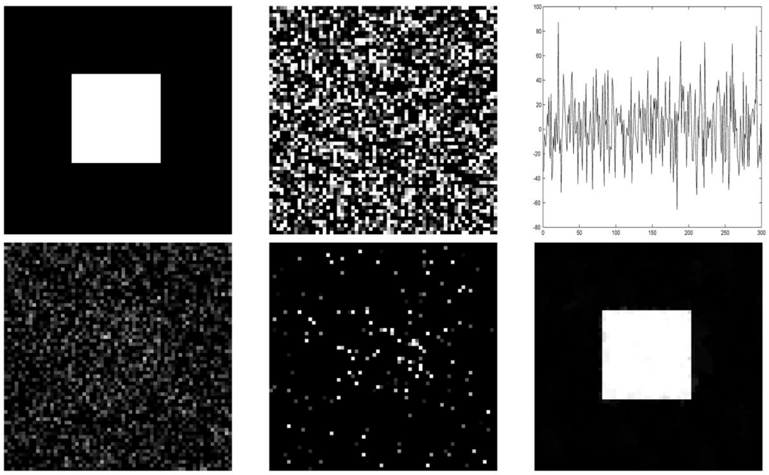

Figure 1.

Results from a simulated data set. The top row includes the true 64 × 64 coefficient image β_0 in the left panel, one realization of a 64 × 64 image predictor X in the middle panel, and the responses Y from $n = 300$ in the right panel. The bottom row includes the estimated coefficient functions obtained from fPCA (left), Lasso (middle), and Total Variation (right).

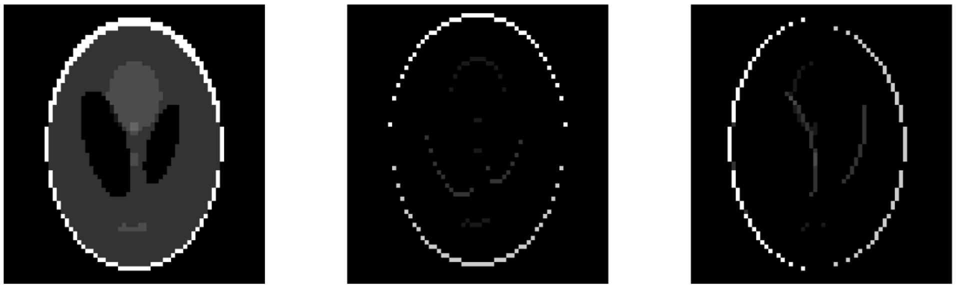

Figure 2.

Left: the Shepp-Logan phantom image; Middle and Right: the two components of the discrete gradient of the phantom image.

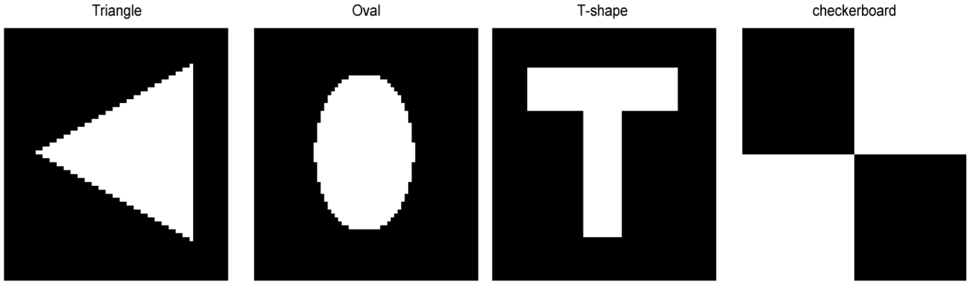

Figure 3.

The true coefficient images used for the simulation study.

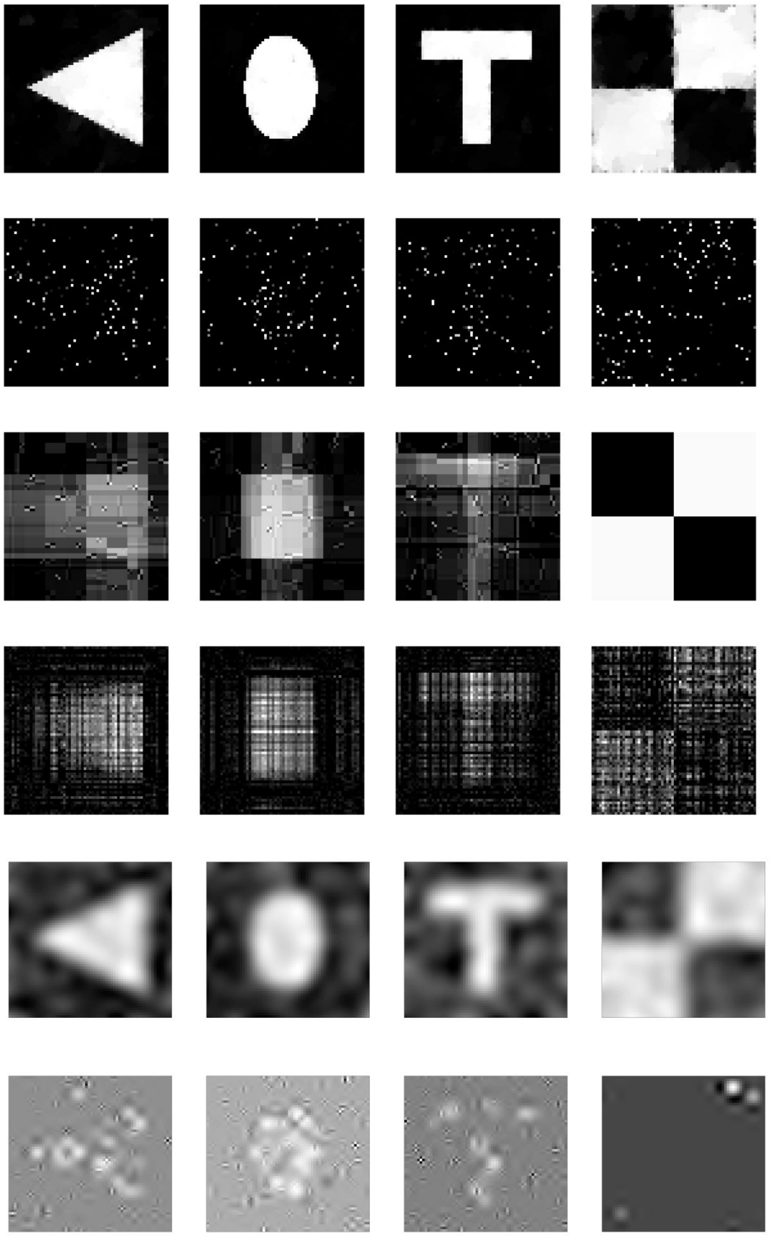

Figure 4.

The estimated coefficient images from six estimation methods when $q = 0$ and $n = 300$: TV (Top row); Lasso (Second row); Lasso-Haar (Third row); Matrix regression (fourth row); FPCR (Fifth row); and WNET(Sixth row).

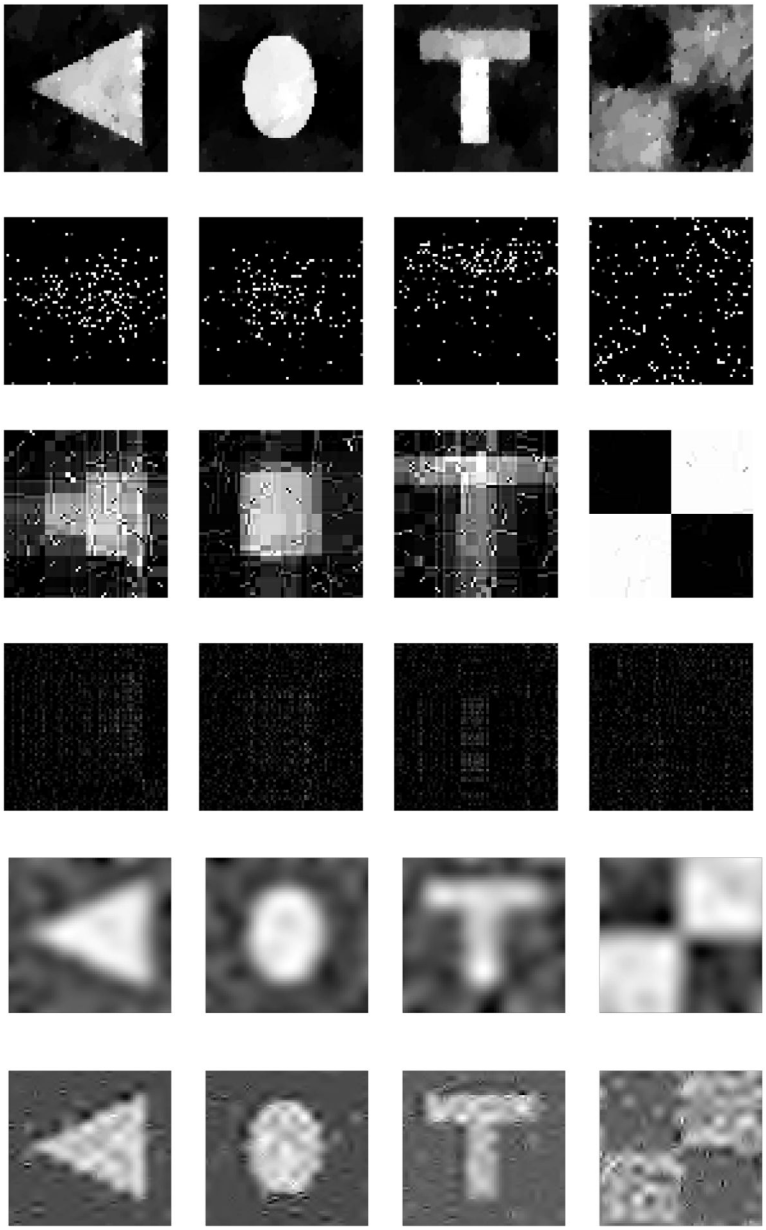

Figure 5.

The estimated coefficient images from six methods when $q = 0.5$ and $n = 300$: TV (Top row); Lasso (Second row); Lasso-Haar (Third row); Matrix regression (fourth row); FPCR (Fifth row); and WNET(Sixth row).

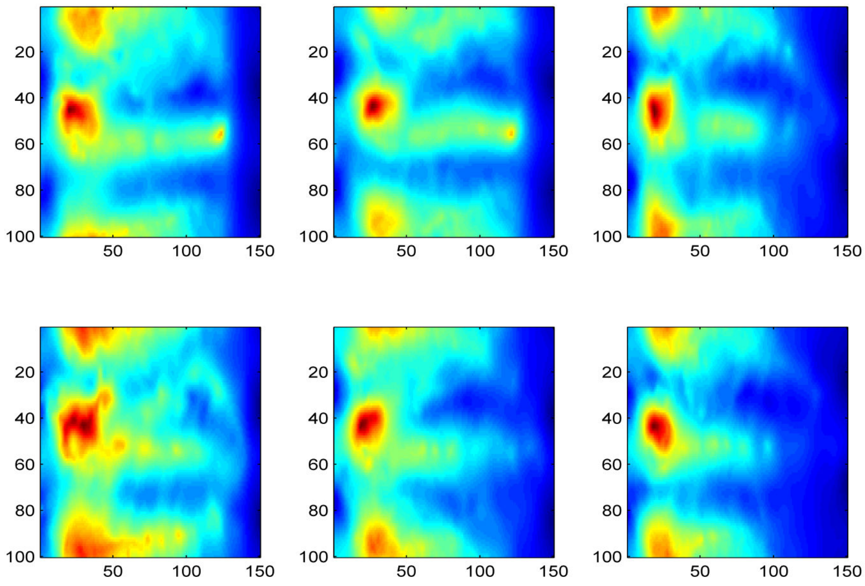

Figure 6.

Observed left hippocampus images.

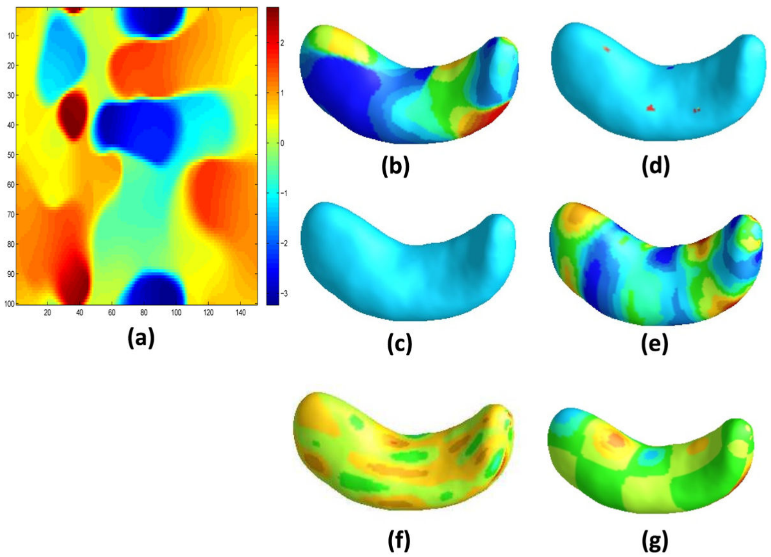

Figure 7.

Estimated coefficient images for hippocampus data based four methods: the 2d-representation of TV estimator (a) and the surface representation of TV estimator (b), Lasso estimator (c), Lasso-wavelet estimator (d), matrix regression estimator (e), FPCR estimator (f), and WNET estimator (g).

Table 1

The RMSPEs of six methods including TV, Lasso, Lasso-Haar, Matrix-Reg, FPCR, and WNET for four different shapes: the numbers in brackets are the corresponding standard errors of those RMSPEs.

	TV			**Lasso**			**Lasso-Haar**		
q	**0**	**0.5**	**1**	**0**	**0.5**	**1**	**0**	**0.5**	**1**
Triangle	3.20 (0.70)	1.53 (0.12)	1.18 (0.08)	34.02 (2.25)	13.12 (0.95)	3.96 (0.32)	18.76 (1.43)	5.08 (0.52)	1.98 (0.17)
Oval	1.69 (0.22)	1.49 (0.13)	1.22 (0.09)	31.16 (2.00)	11.83 (0.95)	3.57 (0.25)	14.99 (1.24)	3.82 (0.28)	1.67 (0.15)
T-shape	2.04 (0.51)	1.47 (0.09)	1.23 (0.10)	30.10 (2.41)	18.81 (0.81)	3.34 (0.28)	19.65 (1.84)	4.62 (0.46)	1.69 (0.16)
Checkerboard	6.27 (1.09)	1.45 (0.11)	1.10 (0.08)	49.00 (3.54)	25.46 (1.65)	9.08 (0.83)	1.43 (0.14)	1.06 (0.08)	1.05 (0.07)
	Matrix-Reg			FPCR			WNET		
q	0	0.5	1	0	0.5	1	0	0.5	1
Triangle	23.13 (1.78)	6.46 (0.55)	4.00 (0.34)	10.89 (0.76)	2.27 (0.18)	1.19 (0.09)	28.42 (2.45)	2.10 (0.17)	1.28 (0.10)
Oval	18.81 (1.58)	5.58 (0.37)	3.41 (0.30)	9.28 (0.68)	2.12 (0.15)	1.19 (0.09)	24.74 (2.40)	2.03 (0.16)	1.26 (0.10)
T-shape	20.97 (1.45)	5.49 (0.38)	3.57 (0.33)	11.69 (0.82)	3.42 (0.26)	1.64 (0.12)	24.41 (2.18)	2.52 (0.21)	1.46 (0.13)
Checkerboard	35.33 (2.89)	12.80 (0.95)	6.35 (0.38)	10.70 (0.83)	3.05 (0.22)	1.14 (0.10)	44.24 (3.23)	5.41 (0.55)	2.03 (0.15)

Table 2

ADNI hippocampus data set: the estimated coefficients of the four scalar covariates and their standard deviations in parentheses.

	intercept	sex	age	behavior score
$\hat{\theta}$	−1.807 (3.186)	−0.533 (0.590)	−0.093 (0.043)	0.869 (0.111)