JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 545, 715–729: Theory and Methods
https://doi.org/10.1080/01621459.2022.2142590

Robust High-Dimensional Regression with Coefficient Thresholding and Its Application to Imaging Data Analysis

Bingyuan Liu[a], Qi Zhang[a], Lingzhou Xue[a], Peter X.-K. Song[b], and Jian Kang[b]

[a]The Pennsylvania State University, University Park, PA; [b]University of Michigan, Ann Arbor, MI

ABSTRACT

It is important to develop statistical techniques to analyze high-dimensional data in the presence of both complex dependence and possible heavy tails and outliers in real-world applications such as imaging data analyses. We propose a new robust high-dimensional regression with coefficient thresholding, in which an efficient nonconvex estimation procedure is proposed through a thresholding function and the robust Huber loss. The proposed regularization method accounts for complex dependence structures in predictors and is robust against heavy tails and outliers in outcomes. Theoretically, we rigorously analyze the landscape of the population and empirical risk functions for the proposed method. The fine landscape enables us to establish both statistical consistency and computational convergence under the high-dimensional setting. We also present an extension to incorporate spatial information into the proposed method. Finite-sample properties of the proposed methods are examined by extensive simulation studies. An application concerns a scalar-on-image regression analysis for an association of psychiatric disorder measured by the general factor of psychopathology with features extracted from the task functional MRI data in the Adolescent Brain Cognitive Development (ABCD) study. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2019
Accepted October 2022

KEYWORDS
Landscape analysis;
Nonconvex optimization;
Scalar-on-image regression;
Thresholding function

1. Introduction

Regression analysis of high-dimensional data has been extensively studied in many research fields over the last three decades. To overcome the high-dimensionality, researchers have proposed a variety of regularization methods to perform variable selection and parameter estimation simultaneously. The ℓ_0 regularization enjoys the oracle risk inequality (Barron, Birgé, and Massart 1999) but it is impractical due to its NP-hard computational complexity. In contrast, the ℓ_1 regularization (Tibshirani 1996) provides an effective convex relaxation of the ℓ_0 regularization and achieves variable selection consistency under the *irrepresentable condition* (Zhao and Yu 2006; Zou 2006; Wainwright 2009). The adaptive ℓ_1 regularization (Zou 2006) and the folded concave regularization (Fan and Li 2001; Zhang 2010) relax the *irrepresentable condition* and improve the estimation and variable selection performance. The folded concave penalized estimation can be implemented through solving a sequence of adaptive ℓ_1 penalized problems and achieves the strong oracle property (Zou and Li 2008; Fan, Xue, and Zou 2014).

Despite these important advances, existing methods, including the (adaptive) ℓ_1 regularization and folded concave regularization, do not work well when predictors are strongly correlated, which is the case especially in scalar-on-image regression analysis (Wang, Zhu, and Initiative 2017; Kang, Reich, and Staicu 2018; He, Xu, and Kang 2018). This article is motivated by the needs of analyzing the n-back working memory task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study (Casey et al. 2018). The task-invoked fMRI imaging measures the blood oxygen level signal that is linked to personal neural activities when performing a specific task. The n-back task is a commonly used approach to making assessment in psychology and cognitive neuroscience with a focus on working memory. One question of interest is to understand the association between the risk of developing psychiatry disorder and features related to functional brain activity. We use the 2-back versus 0-back contrast map statistics derived from the n-back task fMRI data as image predictors. We aim at identifying important imaging biomarkers that are strongly associated with the general factor of psychopathology (GFP) or "p-factor," which is a psychiatric disorder outcome used to evaluate the overall mental health of a subject. In this application, it is expected that the irrepresentable condition can be easily violated by strong dependence among high-dimensional image predictors from fMRI data. To illustrate the presence of strong dependence among image predictors, Figure 1 shows the largest absolute value of correlation coefficients and the number of correlation coefficients that are ≥ 0.8 or ≤ -0.8 between brain regions. Among all pairs between 2518 voxels, there are 151,724 voxel pairs across these regions having a correlation larger than 0.8 (or less than -0.8), and 9,038 voxel pairs with a correlation larger than 0.9 (or less than -0.9). We see that there exists strong dependence among image predictors, so that existing methods

CONTACT Lingzhou Xue lzxue@psu.edu The Pennsylvania State University, University Park, PA.
Bingyuan Liu and Qi Zhang contributed equally to this article.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

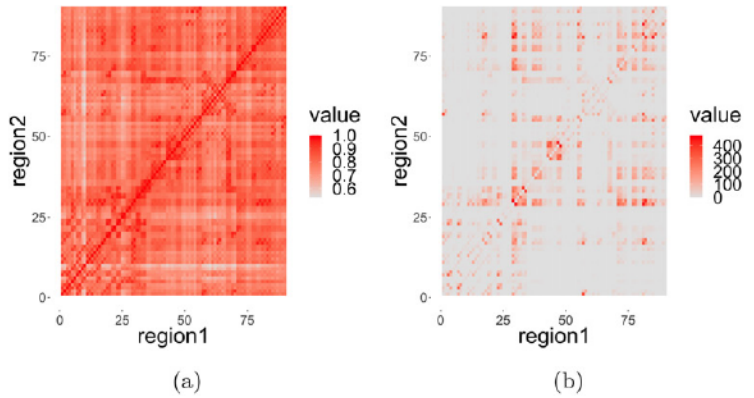


Figure 1. Illustration of the strong dependence structure among image predictors. Panel (a) shows the largest absolute value of correlation coefficients between regions, and Panel (b) shows the number of correlation coefficients that are ≥ 0.8 or ≤ -0.8 between regions.

may not have a satisfactory performance in the scalar-on-image analysis.

To address potential technical challenges in the presence of such strongly correlated predictors, we consider a new approach based on the coefficient thresholding technique. The rationale behind our idea is rooted in attractive properties given by various recently developed thresholding methods, including the *hard-thresholding property* of the ℓ_0 regularization (Fan and Lv 2013) and recovery properties of iterative hard thresholding on badly conditioned problems (Jain, Tewari, and Kar 2014). Especially, Fan and Lv (2013) showed that the global minimizer of the ℓ_0 regularization in the thresholded parameter space enjoys the variable selection consistency. Thus, with proper thresholding of coefficients, it is possible to significantly relax the irrepresentable condition while to address the strong dependence among predictors. Recently, manifested by the potential power of the thresholding strategy, Shi and Kang (2015) and Kang, Reich, and Staicu (2018) studied a new class of Bayesian nonparametric models based on the thresholded Gaussian prior, and Sun et al. (2019) proposed a two-stage hard thresholding regression analysis that applies a hard thresholding function on the initial ℓ_1-penalized estimator.

Beyond the strong dependence among imaging features, there exist two additional challenges in this real application. On the one hand, it is important to integrate the AAL region partition, which provides useful information on the brain structure and function, as grouping information of image predictors to improve the accuracy of imaging feature selection. On the other hand, the outcome variable "p-factor" has a right skewed marginal distribution with heavy tails (and its kurtosis equals to 66). Robustness against outliers occurring from heavy-tailed errors is essential in the scalar-on-image analysis. fMRI indirectly measures neural activity by assessing blood-oxygen-level-dependent signals and its signal-to-noise ratio is often low (Lindquist 2008). Also, due to various limitations of used instruments and quality control in data preprocessing, fMRI data often involves many potential outliers (Poldrack 2012), compromising the stability and reliability of standard regression analyses. The complexity of fMRI techniques limits the capacity of unifying fMRI data preprocessing procedures (Bennett and Miller 2010; Brown and Behrmann 2017) to identify and remove outliers effectively. Standard regression analysis with contaminated data may lead to a high rate of false positives in inference, as shown in many empirical studies (Eklund et al. 2012; Eklund, Nichols, and Knutsson 2016). It is loudly advocated that potential outliers should be taken into account in the study of brain functional connectivity using fMRI data (Rosenberg et al. 2016). These challenges motivate us to design a robust variable selection model against strong dependence among features, heavy tailed distributions and outliers of the response, and accommodate group structure at the same time.

In the current literature of the high-dimensional scalar-on-image regression, Goldsmith, Huang, and Crainiceanu (2014) introduced a single-site Gibbs sampler that incorporates spatial information in a Bayesian regression framework to perform the scalar-on-image regression. Li et al. (2015) introduced a joint Ising and Dirichlet process prior to develop a Bayesian stochastic search variable selection. Wang, Zhu, and Initiative (2017) proposed a generalized regression model in which the image is assumed to belong to the space of bounded total variation incorporating the piece-wise smooth nature of fMRI data. Motivated by these works, in this article we first introduce a new integrated robust regression model with coefficient thresholding and then propose a penalized estimation procedure with provable theoretical guarantees, where the noise distribution is not restricted to be sub-Gaussian. Specifically, we propose to use a smooth thresholding function to approximate the discrete hard thresholding function to tackle the strong dependence of predictors together with the use of the smoothed Huber loss (Charbonnier et al. 1997) to achieve desirable robust estimation. We design a customized composite gradient descent algorithm to efficiently solve the nonconvex and nonsmooth optimization problem. The proposed coefficient thresholding method is capable of incorporating intrinsic group structures of high-dimensional image predictors and dealing with their strong spatial and functional dependencies. Moreover, the proposed method effectively improves robustness and reliability.

The proposed regression with the coefficient thresholding method results in a nonconvex objective function in optimization. In the current literature, it becomes an increasingly important research topic to obtain the statistical and computational guarantees for nonconvex optimization methods. The local linear approximation (LLA) approach (Zou and Li 2008; Fan, Xue, and Zou 2014; Fan et al. 2018) and the Wirtinger flow method

(Candes, Li, and Soltanolkotabi 2015; Cai, Li, and Ma 2016) directly have enabled to analyze the computed local solution. The restricted strong convexity (RSC) condition (Negahban et al. 2012; Negahban and Wainwright 2012; Loh and Wainwright 2013; Jain, Tewari, and Kar 2014; Loh and Wainwright 2017) and the null consistency condition (Zhang and Zhang 2012) were used to prove the uniqueness of the sparse local solution. However, it still remains nontrivial to justify the nice properties of the initial solution for LLA or prove the RSC condition in the presence of strongly dependent predictors. Thus, it is very challenging to study theoretical properties of the proposed robust regression with coefficient thresholding. The nonconvex optimization cannot be directly solved by the LLA approach, and doesn't belong to the family of nonconvex function where RSC condition can be applied. Alternatively, following Mei, Bai, and Montanari (2018), we study the landscape of the proposed method. We prove that the proposed nonconvex loss function has a fine landscape with high probability and also establish the uniform convergence of the directional gradient and restricted Hessian of the empirical risk function to their population counterparts. Thus, under some mild conditions, we can establish key statistical and computational guarantees. Let n be the sample size, p be the dimension of predictors, and s the size of the support set of true parameters. Specifically, we prove that, with high probability, (i) any stationary solution is consistent under the ℓ_2 norm when $n \geq Cs\log p$, where C is a constant; and (ii) the proposed composite gradient descent algorithm attains the desired stationary solution. Both statistical and computational guarantees of the proposed method do not require a specific type of initial solutions.

The rest of this article is organized as follows. Section 2 proposes the robust regression with coefficient thresholding procedure. Section 3 studies theoretical properties of the proposed method, including both statistical guarantees and computational guarantees. Section 4 presents an extension to incorporate the spatial information. Simulation studies are presented in Section 5 and the real application is demonstrated in Section 6. Section 7 includes a few concluding remarks. All the remaining technical details and proofs are given in the supplementary materials.

2. Methodology

In this section, we will first introduce the thresholding function and its motivation in Section 2.1 and then present our proposed robust regression with coefficient thresholding in Section 2.2.

Let $(X_i, Y_i)_{i=1}^n$ be a sample of n independent observations from $(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)^\mathsf{T}$ is a p-dimensional predictor vector and $\mathcal{Y}$ is a scalar response variable. Consider the linear regression $y = X\beta^\star + \varepsilon$, where $y = (Y_1, \ldots, Y_n)^\mathsf{T}$ is the response vector, $X = (x_1, \ldots, x_p)$ is the $n \times p$ deterministic design matrix, $\beta^\star = (\beta_1^\star, \ldots, \beta_p^\star)^\mathsf{T}$ is the coefficient vector, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$ is a random error with mean zero. To be clear, $X_j \in \mathbb{R}^p$ is a row of X while $x_j \in \mathbb{R}^n$ is a column of X. The support of $\beta^\star$ is $S = \{j : \beta_j^\star \neq 0\}$ with cardinality $|S| \ll p$. We aim to recover the true sparse signal $\beta^\star$ given the possibly strong dependence among predictors and heavy-tailed distribution of ε.

2.1. Thresholding Function

The thresholding strategy has been used to deal with strong dependence among predictors by Jain, Tewari, and Kar (2014), Shi and Kang (2015), Kang, Reich, and Staicu (2018), and Sun et al. (2019). Especially in the imaging data analysis, Shi and Kang (2015) proposed the hard-thresholded Gaussian process for selecting important image features, and Kang, Reich, and Staicu (2018) proposed the soft-thresholded Gaussian prior and showed its promising numerical performance. Motivated by these works, we design a thresholding function $g(\cdot) = (g_1(\cdot), \ldots, g_p(\cdot)) : \mathbb{R}^p \to \mathbb{R}^p$ of the coefficient β, to reweight the linear effects $\sum_{j=1}^p x_j\beta_j$ as $\sum_{j=1}^p x_j(\beta_j g_j(\beta_j))$ based on the feature importance under the regression framework. Let $G(\beta) = \beta \circ g(\beta) = (f_1(\cdot), \ldots, f_p(\cdot))$, where $a \circ b$ is the element-wise product between a and b. Then reweighted linear effects $\sum_{j=1}^p x_j(\beta_j g_j(\beta_j))$ can be written using matrix form as $XG(\beta)$. We call this reweight scheme as the coefficient thresholding. The motivation of the coefficient thresholding is as below: suppose we know some oracle information of the true signal $\beta^\star$ before fitting the model, such as the hard thresholding property introduced in the next paragraph, we can design a function g to use the oracle information and adaptively weight different features in the linear effects. In this way, from arbitrary $\beta \in \mathbb{R}^p$, $G(\cdot)$ will first map β to a point $G(\beta)$ better mimic the true signal $\beta^\star$ and then operate on the features x_j. The reweighted linear effects $\sum_{j=1}^p x_j(\beta_j g_j(\beta_j))$ appears in the loss function for goodness of fit. This largely relax the requirement of a "good initial solution" for solving the adaptive ℓ_1 or folded concave penalized problem.

In the following, we only consider the case when $g_1 = \cdots = g_p = g$. Let $\eta^\star = \min_{j\in S}|\beta_j|$ be the minimum true signal strength. The hard thresholding function $I\{|\cdot| \geq \eta^\star\}\}$ would be a good choice for $g(\cdot)$. This is motivated by the best subset selection with the ℓ_0-regularization, which enjoys the oracle risk inequality (Barron, Birgé, and Massart 1999). (Fan and Lv 2013, Proposition 2) further proved the hard-thresholding property of the global solution of ℓ_0-regularized regression problem, that is, each component of the estimator is either 0 or has magnitude larger than some positive threshold. For space consideration, we present the connection between the coefficient thresholding and ℓ_0 regularization in the thresholded parameter space in the supplementary materials.

The discontinuity of $I\{|u| \geq \eta\}$ leads to a challenging optimization problem. Thus, we consider a smooth approximation given by $f_{\tau,\eta}(u) = u \cdot g_{\tau,\eta}(u)$ and $g_{\tau,\eta}(u) = h_\tau(u-\eta) + h_\tau(-u-\eta)$, where $h_\tau(w) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{w}{\tau}\right)$. Here, we use a smooth arctan function to approximate the step function, and the approximation is more accurate when τ gets smaller. As τ goes to 0, $g_{\tau,\eta}(u)$ converges to $g(u)$ pointwisely except at 0. Figure 2 illustrates the smooth approximation of $g_{\tau,\eta}(u)$ to $g(u)$ and $f_{\tau,\eta}(u)$ to $f(u)$. To better control the approximation level and reduce the number of hyperparameters, we fix $\tau/\eta = \varrho$, where constant $\varrho \in \{0.1, 0.01\}$, and the threshold η is left as a tuning parameter whose scale is given in Section 5. Thus, we write $g_{\tau,\eta}$ as g_η and $f_{\tau,\eta}$ as f_η.

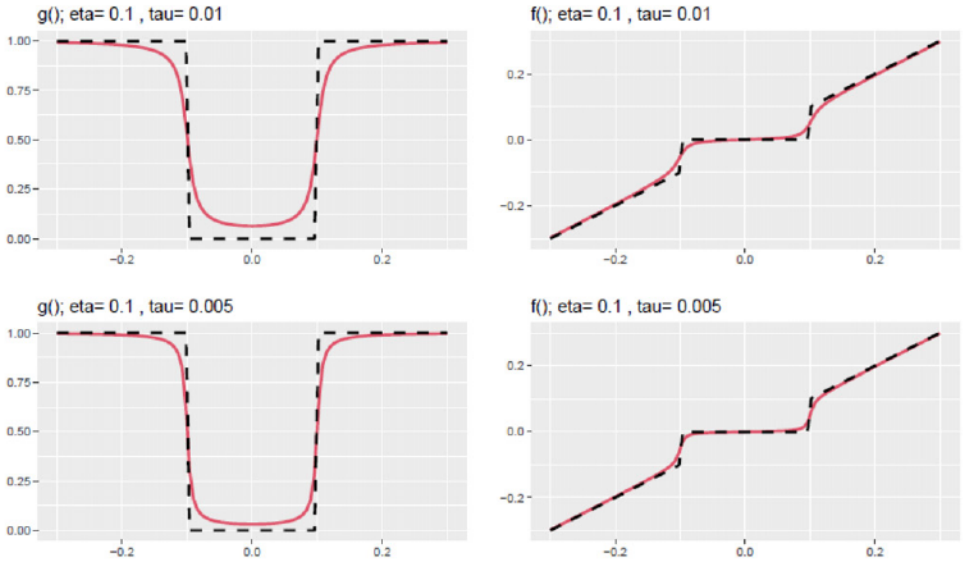


Figure 2. The smooth approximation of $g_{\tau,\eta}(\cdot)$ and $f_{\tau,\eta}(\cdot)$.

2.2. Robust Regression with Coefficient Thresholding

In the lens of robustness, many works have studied the high-dimensional robust regression problem. El Karoui et al. (2013) studied the consistency of regression with a robust loss function such as the least absolute deviation (LAD). In a high-dimensional robust regression, Loh (2017) showed that the use of a robust loss can help achieve the optimal rate of regression coefficient estimation with independent zero-mean error terms. In addition, Loh (2018) showed that by calibrating with a scale estimator in the Huber loss, the regularized robust regression estimator can be further improved. However, existing methods cannot handle the strong dependence among predictors.

To design a robust regression model against strong dependence among features, heavy tailed distributions, and possible outliers, we impose ℓ_1 regularization on the regression coefficients β and propose the following regularized high-dimensional robust regression with coefficient thresholding:

$$\min_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}L(y_i-\sum_{j=1}^{p}x_{ij}\beta_j g_\eta(\beta_j))+\lambda\|\beta\|_1\right\} \quad \text{subject to} \quad \|\beta\|_2\le r, \tag{1}$$

where $L(\cdot)$ is the pseudo-Huber loss (Charbonnier et al. 1997) defined as $L(a)=\omega^2\{\sqrt{1+(a/\omega)^2}-1\}, a\in\mathbb{R}, \omega\in\mathbb{R}$. Note that $L(a)$ provides a smooth approximation of the Huber loss (Huber 1964) and bridges over the l_1 loss and the l_2 loss. In this way, outliers are down-weighted to alleviate potential estimation bias. We note that $L(\cdot)$ has to be differentiable but not necessarily convex in our framework. Other choices, such as Tukey's biweight loss, can also be used to achieve robustness. We shall note that using the thresholding function in Section 2.1 alone cannot avoid overfitting and may not lead to a parsimonious model, especially when features are highly correlated. Consider an extreme case when x_1 and x_2 are identical. Then the thresholding function with threshold $\eta = 0.5$ cannot distinguish whether we should include x_1 or $5x_1-4x_2$ in the model. We will easily obtain an overfitted model if we use thresholding function alone without any regularizations. The regularization in (1) is necessary to pair with the thresholding function.

The framework (1) can be adapted to the group-structure covariates. Suppose the coefficient β can be divided into d separate groups $\beta^1,\ldots,\beta^d$. We consider group Lasso penalty (Yuan and Lin 2006) to leverage known group information in the scalar-on-image analysis. The main model we will analyze in this article is as follows:

$$\min_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}L(y_i-\sum_{j=1}^{p}x_{ij}\beta_j g_\eta(\beta_j))+\lambda\sum_{k=1}^{d}\|\beta^k\|_2\right\} \quad \text{subject to} \quad \|\beta\|_2\le r, \tag{2}$$

where $L(\cdot)$ is still the Pseudo-Huber loss. Other penalties to incorporate group information include sparse group penalty (Simon et al. 2013) and group SCAD penalty (Wang, Chen, and Li 2007).

Following Mei, Bai, and Montanari (2018) and Loh and Wainwright (2017), we assume that the regression coefficients β are bounded in the Euclidean ball $B^p(r)\equiv\{\beta\in\mathbb{R}^p:\ \|\beta\|_2\le r\}$, where r is a constant. As explained by Mei, Bai, and Montanari (2018) and Loh and Wainwright (2017), this assumption is reasonable given the true signal is sparse, and it avoids technical complications. Both (1) and (2) can be efficiently solved by a customized composite gradient descent algorithm with provable guarantees. The details will be presented in Section 3.2. We note that given $\hat\beta$ as the minimizer of (1) or (2), the final estimation should be $G(\hat\beta)$. To further control model sparsity and address the gradient vanishing issue, we use a second step based on hard thresholding, that is,

$$\hat\beta_{\mathrm{RCT}}=\hat\beta\cdot I\{|\hat\beta|\ge\eta\}. \tag{3}$$

This step also helps to address the gradient vanishing issue, which will be explained in Section 3.2. We call $\hat\beta_{\mathrm{RCT}}$ as the RCT (Robust regression with Coefficient Thresholding) estimator.

Remark 1 (Connection with the adaptive Lasso). The RCT estimator simultaneously estimates regression coefficients and adaptive weights to improve the adaptive Lasso (Zou 2006),

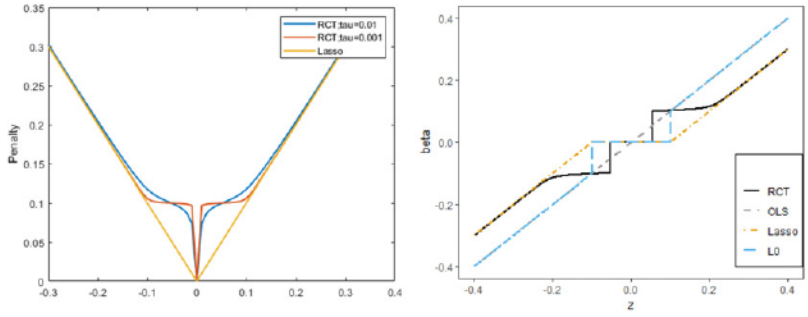


Figure 3. The penalty functions and the univariate solution paths.

whose weights are usually solved from the initial solution or iterated solution. Let $G_\eta = (f_\eta, \ldots, f_\eta)$ and $\xi = G_\eta(\beta)$, where $f_\eta(u) = u \cdot g_\eta(u)$. Consider the ℓ_1 regularized RCT estimation problem. Since G_η is bijective, its inverse $G_\eta^{-1} : \mathbb{R}^p \to \mathbb{R}^p$ exists. If we ignore the constrain, (2) can be equivalently rewritten as

$$\min_{\xi} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(y_i - \sum_{j=1}^{p} x_{ij}\xi_j) + \lambda \sum_{j=1}^{p} \frac{|\xi_j|}{g_\eta(G_\eta^{-1}(\xi)_j)} \right\}. \quad (4)$$

Solving (4) is extremely challenging, since both numerator and denominator of the penalty terms go to zero as $\xi_j \to 0$. Also, solving (4) is not ideal because we still need a good initialization to determine the weights. Especially given the curvature of penalization term (See Figure 3(a)), which is flat and then sharper again as $|\xi_j|$ increases, the solution is very sensitive to the initialization. In our simulation results, we will show that when Lasso fails, adaptive Lasso also fails due to the bad initialization. In the end, the complicated penalty makes it hard to be incorporate the group structure. In comparison, our proposed formulation (2) leads to a nonconvex optimization, which is computationally tractable, not sensitive to initialization, and easily adapted to group penalty structure.

Remark 2 (Comparison with the STGP method). Compared to Kang, Reich, and Staicu (2018), we use a very different approach to incorporate the thresholding function that down weights unimportant variables and achieves sparsity. Our proposed RCT method and its extension in Section 4 are more robust to possible heavy tails and outliers (see the numerical comparison in Table 6 of Section 5). The STGP requires stronger regularity conditions such as the Gaussian error distribution to establish the theoretical properties, and it is unclear about the convergence rate of the posterior computation algorithm for making inferences on the STGP. In addition, the RCT and its extension in Section 4 are more flexible and accurate than the STGP when the nonzero signals are sparsely distributed in a region instead of being spatially connected (see the numerical comparison in Table 5 of Section 5).

Remark 3 (The univariate thresholding rule of RCT). To further illustrate the power of coefficient thresholding in RCT, we consider the univariate solution of penalized least squares using the coefficient thresholding, which is a special case of (1). Assume that each covariate x_j is rescaled to have an L_2-norm $n^{1/2}$. Suppose $\hat{\beta}$ is a global minimizer of (2), then each $\hat{\beta}_j$ minimizes a univariate problem, that is, $\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}} \frac{1}{2}(z - \beta g_\eta(\beta))^2 + \lambda|\beta|$. Then we can get explicit relationship between $\hat{\beta}_j$ and z as $z = \text{sign}(\hat{\beta}_j)\frac{\lambda}{f'_\eta(\hat{\beta}_j)} + f_\eta(\hat{\beta}_j)$. Given this relationship, Figure 3(b) shows the univariate solution path of our solution with Lasso and ℓ_0 penalty. We see that RCT achieves a balance between Lasso and ℓ_0 regularized estimator, and it enjoys the hard thresholding property.

Table 1. Comparison of the RCT and MCP in a small-scale simulation study.

	FPR	FNR
MCP	0.456(0.205)	0.002(0.002)
RCT	0.022(0.085)	0.004(0.003)

In what follows, we use a small simulation study to illustrate the promising performance of the RCT estimator (1) in solving the penalized least squares with strongly dependent predictors, while folded concave penalized methods such as the MCP will perform poorly and tend to include false positives or false negatives of highly correlated covariates. Following He, Xu, and Kang (2018) to mimic the image predictors, we generate predictors with $p = 100$ and $n = 50$ from a Gaussian process covariate structure with high correlation: $\sigma_{ij} = \exp(-\|s_i\|^2 - \|s_j\|^2 - 10\|s_i - s_j\|^2)$, where s_i and s_j are in the rectangle $[-1, 1] \times [-1, 1]$. Let $\beta = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$ and $\epsilon \sim N(0, 3)$. We use MCP and RCT to fit sparse linear regression models. We choose $\tau/\eta = 0.01$ for RCT, and we use cross-validation to choose the tuning parameters for both RCT and MCP. The false positive rate (FPR) and false negative rate (FNR) over 50 replications are reported in Table 1. We see that MCP has significantly higher false positives as it tends to keep lots of correlated covariates with small coefficients, while RCT avoids this issue. Extensive simulation studies will be conducted in Section 5.

3. Theoretical Properties

We present the landscape analysis and asymptotic properties in Section 3.1, and then show the computational guarantee for an efficient composite gradient descent algorithm in Section 3.2.

3.1. Statistical Guarantee

Let $g(u)$ and $f(u)$ be a shorthand of $g_\eta(u)$ and $f_\eta(u)$, respectively. Then $G(\beta) = (f(\beta_1), \ldots, f(\beta_p))$. In the following analysis,

we assume $\tau/\eta = \varrho$, where ϱ is a constant such as 0.01. Let $D_G(\beta) \in \mathbb{R}^{p\times p}$ and $D_G^2(\beta) \in \mathbb{R}^{p\times p\times p}$ be the first two order derivatives of $G(\beta)$, and both $D_G(\beta)$ and $D_G^2(\beta)$ are diagonal. Let $A \preccurlyeq B$ mean that $B - A$ is semi-positive definite. Given $\eta > 0$, G is third continuously differentiable on its domain with the explicit upper and lower bounds for the first two derivatives as in Lemma 1. The proof of Lemma 1 is present in Section B.4 of the supplementary materials.

Lemma 1. (Landscape of thresholding function)

(a) For any $\beta \in B^p(r)$, $\underline{k}_0(\varrho)I_{p\times p} \preccurlyeq D_G(\beta) \preccurlyeq \overline{k}_0(\varrho)I_{p\times p}$, where $\underline{k}_0(\varrho)$ and $\overline{k}_0(\varrho)$ are constants depending on ϱ only.
(b) There exist $\kappa_1(\varrho), \kappa_2(\varrho) > 0$, such that $D_G^2(\beta) \preccurlyeq \overline{m}_0 I_{p\times p\times p}$ when $|\beta_j| \le (\kappa_2(\varrho))\eta$ or $|\beta_j| \ge (1+\kappa_1(\varrho))\eta$ for $j = 1,\dots,p$, where $\overline{m}_0 = c(\varrho)/\eta$. Generally, for any $\beta \in \mathbb{R}^p$, we have $D_G^2(\beta) \preccurlyeq \overline{m}_1 I_{p\times p\times p}$, where $\overline{m}_1 = c_1(\varrho)/\eta$. ($c_1(\varrho) > c(\varrho)$ are constants only depends on ϱ).

We make the following assumptions on the distribution of predictor $\mathcal{X}$, the true parameter $\beta^\star$ and the random error ε.

Assumption 1. (a) The predictor $\mathcal{X} \in \mathbb{R}^p$ is σ^2-sub-Gaussian with mean zero and continuous density $p(\cdot)$, that is, $\mathbb{E}[\mathcal{X}] = 0$ and $\mathbb{E}\{[\exp(\langle u, \mathcal{X}\rangle)]\} \le \exp(\sigma^2\|u\|_2^2/2)$ for any $u \in \mathbb{R}^p$.
(b) The predictor $\mathcal{X}$ spans all directions in $\mathbb{R}^p$, that is, $\mathbb{E}[\mathcal{X}\mathcal{X}^T] \succcurlyeq \gamma\sigma^2 I_{p\times p}$ for some $0 < \gamma < 1$, where σ is specified in (a).
(c) The true parameter $\beta^\star$ has sparsity level $s_0 := \mathrm{supp}(\beta^\star) = o(n)$ and $\|\beta^\star\|_2 \le r$.
(d) The random error ε has a symmetric distribution whose density is strictly positive and decreasing on $(0,\infty)$.

Assumption 1(a) and (b) presents the technical conditions on the predictor. The sub-Gaussian assumption is a commonly used mild condition in high-dimensional regression. Assumption 1(c) imposes the sparsity on the true parameter vector $\beta^\star$. We allow the size of the true support set to diverge at rate $o(n)$. Given the sparsity, it is reasonable to limit our theoretical analysis in the Euclidean ball $B^p(r) := \{\beta \in \mathbb{R}^p, \|\beta\|_2 \le r\}$, which can avoid unnecessary technical complications. Assumption 1(d) allows random error with heavier tails than the standard Gaussian distribution, and it suits for many applications in practice. For example, in our simulation studies (Section 5), the noise is chosen as a mixture of a small variance Gaussian distribution and a large variance Gaussian distribution. Assumption 1(d) can be relaxed to accommodate right skewed errors or discrete errors.

Section 3.1.1 provides the landscape analysis of population risk and empirical risk. Section 3.1.2 further shows the convergence rate of the minimizer of the objective function (2).

3.1.1. The Landscape of Population Risk and Empirical Risk

We analyze the landscape of the population risk function, defined as

$$R(\beta) = \mathbb{E}\Big[L(\mathcal{Y} - \sum_{j=1}^{p}\mathcal{X}_j\beta_j g_{\tau,\eta}(\beta_j))\Big] = \mathbb{E}[L(\mathcal{Y} - \mathcal{X}^{\mathrm{T}}G(\beta))]. \tag{5}$$

We first note that true signal $\beta^\star$ is not a minimizer of $R(\beta)$. Instead, $\tilde{\beta}^\star \in \mathbb{R}^p$ is a minimizer of $R(\beta)$ if $G(\tilde{\beta}^\star) = \beta^\star$ uniquely. Given $\beta^\star \in G(B^p(r))$, the existence and uniqueness of $\tilde{\beta}^\star$ can be guaranteed since the thresholding function G is a bijective between $B^p(r)$ and $G(B^p(r))$. In the sequel, we will study the statistical convergence to the surrogate $\tilde{\beta}^\star$, which shares the same nonzero support with $\beta^\star$. Let $B^p(\beta,\epsilon_0)$ be the ℓ_2 ball centered at β with radius $\epsilon_0 > 0$. Let $\nabla R(\beta)$ be the gradient of $R(\beta)$ and $\nabla^2 R(\beta)$ the Hessian matrix. Let $h(\cdot) = \mathbb{E}_\varepsilon[L'(\cdot+\varepsilon)]$ and $\rho(\cdot) = \inf_{x\le 0}\{h(x)/x\}$. Lemma 2 shows that both $\nabla R(\beta)$ and $\nabla^2 R(\beta)$ are bounded and proves the uniqueness of its stationary point.

Lemma 2. (Landscape of population risk) Under Assumption 1, the population risk function $R(\beta)$ has the following properties:

(a) For any $\beta \in B^p(r)$, $\|\nabla R(\beta)\|_2 \le \overline{L}_0$, where $\overline{L}_0 = \overline{k}_0\omega\sigma$;
(b) For any $\beta \in B^p(r)$, $\langle \beta - \tilde{\beta}^\star, \nabla R(\beta)\rangle \ge T_0\|\beta - \tilde{\beta}^\star\|_2^2$, where $T_0 = \rho(t_0)\gamma\underline{k}_0^2\sigma^2/2$ and $t_0 = (4\overline{k}_0 r\sigma)\sqrt{|\log(\gamma/16)|}$. For any $\beta \in B^p(r)$ such that $\|\beta - \tilde{\beta}^\star\|_2 \ge \epsilon_0$, we have $\|\nabla R(\beta)\|_2 \ge \underline{L}_0$, where $\underline{L}_0 = T_0\epsilon_0$;
(c) For any $\beta \in B^p(r)$, $\lambda_{\max}(\nabla^2 R(\beta)) \le \overline{M}_0$ where $\overline{M}_0 = 4\sigma^2\overline{k}_0^2 + \omega c_1\sigma/\eta$;
(d) There exists an $\epsilon_0 > 0$ such that $\lambda_{\min}(\nabla^2 R(\beta)) \ge \underline{M}_0$ for any $\beta \in B^p(\tilde{\beta}^\star,\epsilon_0)$, where $\underline{M}_0 = \rho'(0)\gamma\underline{k}_0^2\sigma^2/2$,

where $\overline{k}_0$, $\underline{k}_0$, and c_1 are shorthand of $\overline{k}_0(\varrho)$, $\underline{k}_0(\varrho)$, and $c_1(\varrho)$ defined in Lemma 1, depending on ϱ only; σ and γ are constants about the distribution of predictor X, defined in Assumption 1(a) and (b); ω is a constant defined in the Pseudo-Huber loss function $L(\cdot)$; function $\rho(\cdot)$ is a monotone decreasing function depends on the distribution of error ε.

Lemma 2(b) indicates that no stationary point of $R(\beta)$ exists except $\tilde{\beta}^\star$. Lemma 2(d) indicates that Hessian matrix $\nabla^2 R(\beta)$ is strictly positive definite inside the ball $B^p(\tilde{\beta}^\star,\epsilon_0)$. Combine results in (b) and (d), we further conclude that $\beta^\star$ is the unique minimizer.

We then consider the empirical risk function:

$$\widehat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} L\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j g_{\tau,\eta}(\beta_j)\Big). \tag{6}$$

Let $\nabla\widehat{R}_n(\beta)$ and $\nabla^2\widehat{R}_n(\beta)$ be the gradient and Hessian of $\widehat{R}_n(\beta)$. We first establish the uniform convergence from $\nabla\widehat{R}_n(\beta)$ to $\nabla R(\beta)$ and from $\nabla^2\widehat{R}_n(\beta)$ to $R(\beta)$. Thus, fine landscape of the population risk $R(\beta)$ in Lemma 2, such as existence of unique minimizer, can be transferred to the empirical risk $\widehat{R}_n(\beta)$. The landscape properties of empirical risk $\widehat{R}_n(\beta)$ is summarized in Lemma 3.

Lemma 3. (Landscape of empirical risk) Under Assumption 1, for any $\delta > 0, \epsilon_0 > 0$, there exists a constant $C_1 = C_0\sigma^4\log(\overline{m}_1\overline{k}_0 r\sigma/\delta)(\underline{L}_0\epsilon_0 \vee \underline{M}_0)^2$, such that, if $n \ge C_1 p\log p$, the following properties hold with probability at least $1-\delta$:

(a) For any $\beta \in B^p(r)$, $\|\nabla\widehat{R}_n(\beta)\|_2 \le \overline{L}_0/2$.
(b) For any $\beta \in B^p(r)$ such that $\|\beta - \tilde{\beta}^\star\|_2 \ge \epsilon_0/2$, $\langle \beta - \tilde{\beta}^\star, \nabla\widehat{R}_n(\beta)\rangle \ge \epsilon_0\underline{L}_0\|\beta - \tilde{\beta}^\star\|_2/4$.
(c) For any $\beta \in B^p(r)$, $\lambda_{\max}(\nabla^2\widehat{R}_n(\beta)) \le \overline{M}_0/2$.

(d) For any $\beta \in B^p(\tilde{\beta}^\star, \epsilon_0)$, $\lambda_{\min}(\nabla^2 \widehat{R}_n(\beta)) \geq \underline{M}_0/2$.
(e) The empirical risk function $\widehat{R}_n(\beta)$ has a unique minimizer $\hat{\beta}_n$ such that

$$\|\hat{\beta}_n - \tilde{\beta}^\star\|_2 \leq C_2\sqrt{p \log n/n},$$

where $C_2 = 4C_0\delta \log(\overline{m}_1\overline{k}_0 r\sigma/\delta)/\underline{M}_0$,

where $\overline{L}_0$, $\underline{L}_0$, $\overline{M}_0$, and $\underline{M}_0$ are specified in Lemma 2.

Lemma 3(e) implies the consistency of the unique minimizer of the empiricial risk function when the dimension p diverges with the sample size n under the low-dimensional setting with $n \geq C_1 p \log p$. In the sequel, we further establish the consistency of the RCT estimator when the dimension p can be much larger than the sample size n.

3.1.2. Consistency of RCT Estimator

In this section, we focus on the group lasso penalty case. Let $\beta^\mathsf{T} = ((\beta^1)^\mathsf{T}, \ldots, (\beta^d)^\mathsf{T}) \in \mathbb{R}^{\otimes_{j=1}^d l_j}$, where β_j is a subvector with length l_j, corresponding to the features in group j. For any support index set $S \subset \{1, \ldots, d\}$, let $\beta_S = ((\beta_j)^\mathsf{T}, \ldots)^\mathsf{T}$, where $j \in S$. Let S_0 be the support of $\tilde{\beta}^\star$ on the group index and $d_0 = |S_0|$. Let $d_c = d - d_0$. Let l_j be the length of subvector β_j for $j = 1, \ldots, d$. Let $l_{S_0} = \max\{l_j, j \in S_0\}$, $l_{S_0^c} = \max\{l_j, j \in S_0^c\}$, and $l_{\max} = \max\{l_j, j = 1, \ldots, d\}$.

Given the landscape analysis in the previous section, we will establish the consistency of the RCT estimator when p is at the nearly exponential order of n. To simplify the asymptotic result, as in Mei, Bai, and Montanari (2018), we make the additional assumption on the feature X as follows.

Assumption 2. The feature vector $\mathcal{X}$ is bounded, that is, there exist a constant M, such that, $\|\mathcal{X}\|_\infty < M\sigma$ almost surely.

Assumption 2 is stronger than Assumption 1(a). However, as noted in Mei, Bai, and Montanari (2018), for independent sub-Gaussian data $\{X_i\}_{i=1}^n$, we have $\sup_i \|X_i\|_\infty < C\sqrt{\log(np)}\sigma$ with high probability. Thus, Theorem 1 can also be established for sub Gaussian features with an additional $\sqrt{\log(np)}$ factor in the error bound. To prove Theorem 1, we first show that the sample directional gradient and restricted Hessian converges uniformly to their population counterparts (Lemma S.6 in supplementary materials). The proof of Theorem 1 is attached in supplementary material B.8.

Theorem 1. Under Assumptions 1 and 2, for any $\delta > 0$, there exist constants C_n, C_λ and C that depend on $(r, \sigma^2, \gamma, \varrho, \eta, M, l)$ but independent of n, p, s_0, such that as $n \geq C_n s_0 \log p$ and $\lambda_n \geq C_\lambda\sqrt{(\log p)/n}$, then with probability at least $1-\delta$, any stationary point $\hat{\beta}$ of group-regularized risk minimization (2) satisfies

$$\|\hat{\beta} - \tilde{\beta}^\star\|_2 \leq C\sqrt{(s_0 \log p)/n + s_0\lambda_n^2}.$$

3.2. Computational Guarantee

Gradient descent algorithms do not work since the objective function (2) is not differentiable at zero. We consider the composite gradient descent algorithm (Nesterov 2013), which is computationally efficient for solving the nonsmooth nonconvex optimization and enjoys the convergence property. Specifically, solving the proposed RCT problem via composite gradient descent algorithm contains two key steps at each iteration: the gradient descent step and the ℓ_2-ball projection step.

In the first step, we perform gradient descent. Given the previous iterated solution $\hat{\beta}^{(k)}$, with the step size h, we need to solve the following subproblem:

$$\min_\beta \left\{ \frac{1}{2}\|\beta - (\hat{\beta}^{(k)} - \frac{1}{h}\nabla\hat{R}(\hat{\beta}^{(k)}))\|_2^2 + \frac{\lambda}{h}\sum_{j=1}^d \|\beta^j\|_2 \right\}. \quad (7)$$

Note that (7) has a closed-form solution through the following group-wise soft thresholding operator: $S_{\lambda/h}(\xi) = \frac{\xi^j}{\|\xi^j\|_2} \circ (\|\xi\|_2 - \lambda/h)_+$, where $\xi = ((\xi^1)^\mathsf{T}, \ldots, (\xi^d)^\mathsf{T})^\mathsf{T} \in \mathbb{R}^p$ shares the same group structure as β, and $\circ$ denotes the Hadamard product. Thus, the gradient descent step can be solved as

$$\tilde{\beta}^{(k+1)} = S_{\lambda/h}(\hat{\beta}^{(k)} - h\nabla(\widehat{R}_n(\hat{\beta}^{(k)}))). \quad (8)$$

In the second step, we project $\tilde{\beta}^{(k+1)}$ onto the ℓ_2-ball by

$$\pi_r(\tilde{\beta}^{(k+1)}) = \frac{\min\{\|\tilde{\beta}^{(k+1)}\|_2, r\}}{\|\tilde{\beta}^{(k+1)}\|_2}\tilde{\beta}^{(k+1)}. \quad (9)$$

After solving (8) and (9) until convergence, we apply a hard thresholding on the solution to get the final RCT estimator, that is, $\hat{\beta}_{\text{RCT}} = \hat{\beta} \cdot I\{|\hat{\beta}| \geq \eta\}$. Note that $G(\cdot)$ is a smooth approximation of the hard thresholding function, so this additional step results in a sparse estimator $\hat{\beta}$ close to $G(\hat{\beta})$. Moreover, this hard thresholding step helps to address the gradient vanishing issue, which we will illustrate as follows. To begin with, we derive an upper bound of the partial derivative of the loss function at one observation with respect to β_j, that is,

$$\left|\frac{\partial L(y_i, \beta_j)}{\partial \beta_j}\right| = \left|L'(y_i - \sum_{j=1}^p x_{ij}\beta_j g_\eta(\beta_j))\frac{\partial \beta_j g_\eta(\beta_j)}{\partial \beta_j}\right|$$
$$\leq |g(\beta_j) + \beta_j g'(\beta_j)|.$$

Figure 4 plots the upper bound for $\beta \in (0, 1)$ with $\eta = 0.5$ and $\tau = 0.005$. We see that the derivative almost vanishes

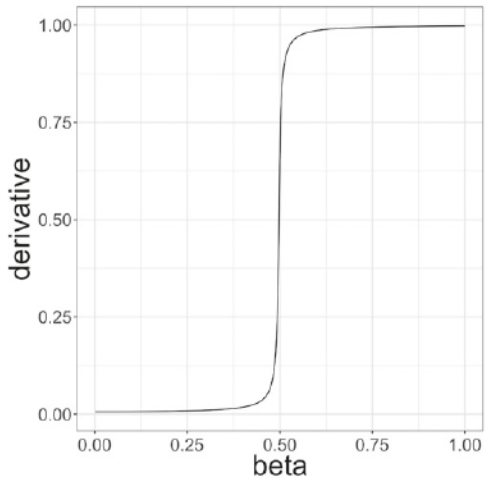


Figure 4. Illustration of the gradient vanishing issue for $\beta \in (0, 1)$.

Input: $\beta^{(0)} \in B^p(r)$, step size h, penalization parameter λ, thresholding parameter η, and predetermined hyperparameter τ, r
for $k = 0, 1, 2, \ldots$ *until convergence* **do**
  $\tilde{\beta}^{(k+1)} = S_{\lambda/h}(\hat{\beta}^{(k)} - h\nabla(\widehat{R}_n(\hat{\beta}^{(k)})))$
  $\hat{\beta}^{(k+1)} = \pi_r(\tilde{\beta}^{(k+1)})$
end
Record the solution as $\hat{\beta}$.
Output: $\hat{\beta}_{\text{RCT}} = \hat{\beta} \cdot I\{|\hat{\beta}| \geq \eta\}$.

Algorithm 1: The composite gradient descent algorithm for solving the RCT estimator.

when $|\beta| < \eta$. That means, after the updated solution enters the threshold (i.e., $|\beta_j| < \eta$), vanishing gradient will prevent the solution approaching 0. Then the soft-thresholding (8) is not enough to threshold the noise signals to 0. Such gradient vanishing is a common issue in optimization when the objective function is nonconvex (Hochreiter 1998). In our context this hard thresholding step addresses this issue, as gradient vanishing only happens when $|\beta| < \eta$.

The proposed algorithm can be summarized as Algorithm 1. At each iteration, the subproblem can be solved with a closed-form solution, and the computational complexity is on the quadratic order of dimension p. The algorithmic convergence rate is presented in the following proposition.

Proposition 1. Let $\hat{\beta}^{(k)}$ be the kth iterated solution of Algorithm 1. There exist constants c_h and C, independent of (n, p, s_0), such that when $h < c_h$, there exists $k < C\epsilon^{-2}$ and subgradient $u((\hat{\beta}^{(k)})^j) \in \partial\|(\hat{\beta}^{(k)})^j\|_2$, such that

$$\|\nabla\widehat{R}_n(\hat{\beta}^{(k)}) + \lambda\sum_{j=1}^{d} u((\hat{\beta}^{(k)})^j)\|_2 \leq \epsilon,$$

where $\partial\|\beta^j\|_2$ denotes the sub-differential of the group penalty function.

Proposition 1 provides a theoretical justification of the algorithmic convergence rate. The proposed algorithm always converges to an approximate stationary solution (a.k.a. ϵ-stationary solution) at finite sample sizes. In other words, after $O(1/\epsilon^2)$ iterations, the ℓ_2 norm of the subgradient of the objective function is bounded by ϵ when the sample size is finite. When k increases, the proposed algorithm will find the stationary solution that satisfies the subgradient optimality condition as $\epsilon \to 0$. To better visualize the convergence of the algorithm, we provide the convergence plots and computational cost of the proposed algorithm in Section C of the supplementary materials. From both theoretical and practical aspects, the proposed algorithm is computationally efficient and achieves the desired computational guarantee. Given the nice empirical gradient structure proved in Theorem 1, we further prove the linear convergence rate given that the solution is sparse in Proposition 3 in Section C of the supplementary materials.

4. Extension with the Spatial Information

This section extends the methodology and applicability of the RCT estimator to incorporate the possibly available spatial information among predictors, especially for scalar-on-image regression. In practice, the scalar-on-image regression model has a large number of pixels or voxels as the predictors, but only a few are significantly associated with the response. When a pixel or voxel is selected as a significant one, its neighbors on the image usually have similar effects. The spatial information is commonly available and used in image data analysis such as the STGP (Kang, Reich, and Staicu 2018).

Denote the neighbor index set of β_j by A_j, and denote the subvector of β on A_j by β_{A_j}. Let $n_j = |A_j|$. Let $\bar{\beta}_{A_j}$ be the average of all neighbor signals of β_j, that is, $\bar{\beta}_{A_j} = \sum_{k\in A_j}\beta_k/n_j$. For a two-dimensional image, we define the neighbors of a predictor as the ones whose either row index or column index differs by 1 (i.e., $n_j = 4$). Recall that we employ the coefficient thresholding on each coordinate in the proposed RCT method (2). We now modify it by thresholding over the average between coefficient β_j and its average neighbor effect $\bar{\beta}_{A_j}$, that is,

$$\tilde{g}_{\tau,\eta}(\beta_j) = h_\tau((\beta_j + \bar{\beta}_{A_j})/2 - \eta) + h_\tau((-\beta_j - \bar{\beta}_{A_j})/2 - \eta).$$

Intuitively, when the neighbors of x_j have significant effects on the response, $|\bar{\beta}_{A_j}|$ tends to be above the threshold η and helps to keep x_j in the model, because the effects usually change smoothly across different locations and have the same sign within a neighbor in the scalar-on-image regression setting.

After incorporating the spatial information to the coefficient thresholding, we then solve the following penalized problem:

$$\min_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}L(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\tilde{g}_{\tau,\eta}(\beta_j)) + \lambda_n\sum_{k=1}^{d}(\|\beta^k\|_2)\right\}$$
$$\text{subject to} \quad \|\beta\|_2 \leq r. \quad (10)$$

After incorporating the spatial information, it is challenging to analyze the landscape of the risk function in (10), and there is no longer any guarantee of a unique stationary solution with high probability. To handle the issue of multiple local solutions, we propose a stochastic composite gradient descent algorithm that also enjoys a convergence guarantee (see Proposition 2). For the kth iteration, define $x_1^{b_k}, \ldots, x_t^{b_k}$ as the randomly split batches with the given batch size and $R_q^{b_k}$ as the empirical loss function calculated on the qth batch in the kth iteration. The proposed stochastic composite gradient descent algorithm is proceeded as follows.

Let $L(\beta) = R(\beta) + \lambda\sum_{k=1}^{d}\|\beta_k\|_2$ and $L_n(\beta) = \widehat{R}_n(\beta) + \lambda\sum_{k=1}^{d}\|\beta_k\|_2$. Let $L^* = \min_{\beta\in B^p(r)}L(\beta)$. We now establish the following convergence guarantee for the above stochastic algorithm based on (Ghadimi, Lan, and Zhang 2016, Theorem 1).

Proposition 2. Suppose that the step size $\{h_k\}$ in the Algorithm 2 are chosen such that $0 < h_k \leq 2\alpha/\overline{M}_0$ with $h_k < 2\alpha/\overline{M}_0$ for at least one k, where $\overline{M}_0$ is defined in Lemma 2(c). Then, we have

$$\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\| \leq \frac{\overline{M}_0 D^2}{\sum_{j=1}^{k}(\alpha h_k - \overline{M}_0 h_k^2/2)},$$

Input: $\beta^{(0)} \in B^p(r)$, step size h_k, penalization parameter λ, thresholding parameter η, batch size b, and predetermined hyperparameter τ, r.
for $k = 0, 1, 2, \ldots$ *until convergence* **do**
  Split data into batches $x_1^{b_k}, \ldots, x_t^{b_k}$
  for $q = 0, 1, 2, \ldots, t$ **do**
    $\tilde{\beta}^{(k;q+1)} = S_{\lambda/h_k}(\hat{\beta}^{(k;q)} - h_k \nabla(\widehat{R}_q^{b_k}(\hat{\beta}^{(k;q)})))$
    $\hat{\beta}^{(k;q+1)} = \pi_r(\tilde{\beta}^{(k;q+1)})$
  end
  $\hat{\beta}^{(k+1;0)} = \hat{\beta}^{(k;t+1)}$
end
Output: $\hat{\beta}_{\mathrm{RCT}} = \hat{\beta} \cdot I\{|\hat{\beta}| \geq \eta\}$.

Algorithm 2: The proposed stochastic composite gradient descent algorithm.

where $D = \sqrt{(L(\beta^{(0)}) - L^*/\overline{M}_0)}$. If we take $h_k = \alpha/\overline{M}_0$ for all k, then we have

$$\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\| \leq \frac{2\overline{M}_0^2 D^2}{\alpha^2 k}.$$

In the simulation studies of Section 5, we will show that the generalized RCT method (10) and the stochastic algorithm are promising for the scalar-on-image regression analysis.

5. Simulation Studies

This section examines the finite-sample properties of the RCT estimator and its extension in simulation studies. Section 5.1 examines not only the linear regression settings but also the scalar-on-image regression settings without spatial information. Section 5.2 examines the scalar-on-image regression settings when spatial information is available and can be incorporated into the model.

Specifically, we compare the proposed RCT estimators with the Lasso, Adaptive Lasso (denoted by AdaLasso), SCAD and MCP penalized estimators in four different linear regression models (Models 1–4) and with the Lasso, Group Lasso (denoted by GLasso), and Sparse Group Lasso (denoted by SGL), and the STGP in six Gaussian process regression models (Models 5–10), which mimic the scalar-on-image regression analysis. We denote by RCT and STGP when the spatial information is not used, and by RCT (info)/STGP (info) when the spatial information is used. We implement the Lasso and the Adaptive lasso estimators using R package "glmnet," and we use the Lasso as the initial estimate for the adaptive Lasso. The GLasso is implemented using the method in (Yang 2015). The SCAD and MCP estimators are implemented using R package "ncvreg," and we also verify that the estimation results are consistent with R package "Picasso."

The estimation accuracy is measured by the root-mean-square error (RMSE, that is $\|\hat{\beta} - \beta^\star\|_2$) and the variable selection performance is measured by both false positive rate (FPR) and false negative rate (FNR). Specifically, let FPR = FP/(FP + TN), and FNR = FN/(FN + TP), where TN, TP, FP, and FN represent the numbers of true negative, true positive, false positive and false negative, respectively. Each measure is computed as the average over 50 independent replications.

Before proceeding, we explain the selection of parameters for the proposed methods and algorithms. The penalization parameter λ controls the scale of penalization, and we choose λ by 3-folded cross-validation based on the L_1 prediction error. We set ω as 1 in the pseudo-Huber loss. The radius r of the feasible region can be chosen as a large constant such that the estimator lies in the interior of $B^p(r)$. In all simulations, we set $r = 20$. Next, η and τ are parameters in the thresholding function $g_{\tau,\eta}(\beta)$. τ/η controls the difference between $g_{\tau,\eta}(\beta)$ and ideal hard thresholding function $I(|\cdot| \geq \eta)$. A more rigorous result on how solution depends on τ is shown in Proposition S.1 in the supplementary material. We explore different ways of choosing η and τ in our analysis to verify that results are robust. For the choice of η, in simulation study, we choose η by cross-validation, together with λ. Simulation studies show that any quantiles between 10% and 50% of the Lasso estimator's nonzero coefficients for η give comparable results. In real data study, we let η be the 30% lower-quantile of the absolute value of the nonzero coefficients estimated from the Lasso. Given η, we fix the ratio $\tau/\eta = 0.1$ in Models 5–8, and choose $\tau = 0.02$ in Models 1–4. We explore these two choices to verify that our results are robust against the choice of τ. Lastly, the step size h is chosen to be small enough such that the algorithm doesn't encounter overflow issues. We set h to be 0.01 after we normalize the feature marginally. No nonconvergence issue happens in all of our settings. h can also be chosen according to an acceleration process in Nesterov (2013) to achieve faster convergence.

5.1. Simulation without Spatial Information

First, we consider the linear regression model $\mathcal{Y} = \mathcal{X}^{\mathrm{T}}\beta + \varepsilon$. We generate $\mathcal{X} \sim N(0, \Sigma)$, with the following four different correlation structures for $\Sigma = (\sigma_{ij})_{p\times p}$:

$$\begin{aligned}
&\text{Model 1}: \sigma_{ij} = 0.5^{|i-j|}, && \text{AR1(0.5)}\\
&\text{Model 2}: \sigma_{ij} = 0.7^{|i-j|}, && \text{AR1(0.7)}\\
&\text{Model 3}: \sigma_{ij} = 0.4 + 0.6I(i=j), && \text{CS(0.4)}\\
&\text{Model 4}: \sigma_{ij} = 0.6 + 0.4I(i=j), && \text{CS(0.6)}.
\end{aligned}$$

Models 1–2 have autoregressive (AR) correlation structures, in which the irrepresentable condition for Lasso holds for Model 1 but fails for Models 2. Models 3–4 have the compound symmetry (CS) correlation structures and the irrepresentable condition for Lasso fails in both two models.

We consider $\varepsilon \sim 0.9N(0, \sigma_1^2) + 0.1N(0, \sigma_2^2)$, where σ_2^2 is set much larger than σ_1^2. For Models 1 and 2, consider setting (a) $\sigma_2^2 = 10$, $\sigma_1^2 = 1$ and (b) $\sigma_2^2 = 10$, $\sigma_1^2 = 2$. For Models 3 and 4, consider setting (a) $\sigma_2^2 = 3$, $\sigma_1^2 = 0.1$ and (b) $\sigma_2^2 = 3$, $\sigma_1^2 = 0.3$. When the ratio σ_2^2/σ_1^2 increase, ε has a heavier tail and more extreme values. For all the models, we choose $n = 100$, $p = 2000$ to create a high-dimensional regime and let true signal $\beta^\star = (u, 0_{1980})$, where $u \in \mathbb{R}^{20}$ and $u_i \sim \mathrm{unif}(0.5, 1)$ iid for $i = 1, \ldots, 20$.

Tables 1 and 2 summarize the simulation results for Models 1–4. We have the following observations from these tables. First, in Models 1–2, our RCT estimator outperforms Lasso, adaptive

Table 2. Estimation and selection accuracy of different methods for Models 1–4.

	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss
	Model (1a)			Model (2a)			Model (3a)			Model (4a)		
Lasso	0.021	0.199	3.198	0.014	0.153	3.035	0.040	0.337	4.075	0.041	0.374	4.144
	(0.009)	(0.130)	(0.475)	(0.008)	(0.091)	(0.391)	(0.004)	(0.135)	(0.501)	(0.004)	(0.135)	(0.422)
AdaLasso	0.020	0.212	3.787	0.014	0.156	3.739	0.033	0.369	4.035	0.032	0.443	4.512
	(0.008)	(0.143)	(0.821)	(0.008)	(0.111)	(0.815)	(0.003)	(0.138)	(0.626)	(0.003)	(0.135)	(0.763)
SCAD	0.007	0.422	4.148	0.010	0.575	5.979	0.021	0.736	5.849	0.020	0.745	5.711
	(0.006)	(0.137)	(0.950)	(0.006)	(0.107)	(1.054)	(0.007)	(0.147)	(1.238)	(0.009)	(0.149)	(1.151)
MCP	0.003	0.625	4.709	0.003	0.694	6.201	0.008	0.868	6.763	0.007	0.921	7.348
	(0.002)	(0.065)	(0.537)	(0.002)	(0.049)	(0.536)	(0.003)	(0.084)	(0.558)	(0.003)	(0.064)	(0.510)
RCT	0.010	0.177	2.860	0.002	0.018	1.466	0.061	0.215	3.982	0.066	0.253	4.093
	(0.008)	(0.117)	(0.888)	(0.002)	(0.035)	(0.502)	(0.007)	(0.089)	(0.265)	(0.009)	(0.098)	(0.314)
	Model (1b)			Model (2b)			Model (3b)			Model (4b)		
Lasso	0.019	0.243	3.446	0.014	0.187	3.273	0.040	0.351	4.092	0.041	0.364	4.124
	(0.010)	(0.140)	(0.372)	(0.008)	(0.087)	(0.422)	(0.004)	(0.135)	(0.448)	(0.003)	(0.081)	(0.235)
AdaLasso	0.018	0.256	4.134	0.013	0.199	4.062	0.033	0.377	4.083	0.033	0.459	4.553
	(0.008)	(0.130)	(0.657)	(0.008)	(0.107)	(0.726)	(0.003)	(0.139)	(0.561)	(0.003)	(0.145)	(0.819)
SCAD	0.008	0.443	4.233	0.009	0.566	5.726	0.022	0.719	5.563	0.020	0.744	5.802
	(0.005)	(0.140)	(0.998)	(0.007)	(0.153)	(1.393)	(0.022)	(0.135)	(1.178)	(0.008)	(0.157)	(1.229)
MCP	0.003	0.636	4.771	0.003	0.708	6.386	0.008	0.868	6.738	0.008	0.900	7.146
	(0.002)	(0.069)	(0.627)	(0.002)	(0.058)	(0.667)	(0.003)	(0.079)	(0.600)	(0.004)	(0.080)	(0.719)
RCT	0.018	0.242	3.879	0.007	0.084	2.939	0.060	0.226	4.019	0.066	0.275	4.138
	(0.016)	(0.163)	(0.735)	(0.005)	(0.127)	(0.755)	(0.008)	(0.090)	(0.249)	(0.007)	(0.089)	(0.255)

Lasso (AdaLasso) and nonconvex estimators (SCAD and MCP) more obviously as the auto correlation increases. Nonconvex estimators do not work well on all these settings, since they tend to penalize the model too much, and result in much higher false negative rates. The Lasso estimator misses many true signals due to the highly correlated predictors, leading to bad performance of adaptive Lasso given the Lasso initials. Especially in Model 2 when the auto correlation is high, our RCT estimator has much smaller FPR and FNR compared to Lasso-type methods. Second, in case (a) with more outliers, our estimator achieves a relatively better performance than other estimators thanks to the use of the Pseudo-Huber loss. Third, in more challenging Models 3 and 4, our estimator is able to identify more true predictors and well controls false positives. In summary, RCT estimator is more favorable in high-dimensional regression setting, especially when the predictors are highly correlated. Nonconvex estimators, such as the SCAD and MCP, do not work well in our simulation settings when the dimension is very high compared to the sample size and dependence among predictors is very strong. In the existing literature, Zhang (2010) considered $n = 300$ and $p = 2000$, with features being generated independently; Fan, Xue, and Zou (2014) considered $n = 100$ and $p = 1000$ with the AR1(0.5) covariance matrix; and Loh and Wainwright (2017) used settings with $p = 512$, $n > 100$ and a family of spiked identity covariance. In comparison to those simulation studies, our settings are more challenging, and the SCAD and MCP perform poorly.

Next, we present simulation results on the scalar-on-image regression settings. Consider $\mathcal{Y} = \mathcal{X}^{\mathrm{T}}\beta + \varepsilon$. We generated feature $\mathcal{X}$ with a two-dimensional image structure, whose region was generated by Gaussian processes with mean zero and stationary covariance function $\mathrm{cov}(\mathcal{X}(s), \mathcal{X}(s')) = \kappa(s, s')$ for some pre-specified covariance function κ. Set $n = 500$ and $p = 50 \times 50$, and consider the following covariance structures:

Model 5 (GP1(10)) :

$$\kappa(s_i, s_j) = \exp(-\|s_i\|^2 - \|s_j\|^2 - 10\|s_i - s_j\|^2),$$

Model 6 (GP1(5)) :

$$\kappa(s_i, s_j) = \exp(-\|s_i\|^2 - \|s_j\|^2 - 5\|s_i - s_j\|^2),$$

where $\{s_i \in \mathbb{R}^2, i = 1, \ldots, 2500\}$ are two-dimensional grid equally spaced over the rectangle $[-1, 1] \times [-1, 1]$. Similar to the simulation settings in He, Xu, and Kang (2018), the coefficients with nonzero effects are located on a circle with radius 0.1 on the graph and the values of the nonzero effects are set as 1. The errors follow the same mixture model as in Models 1–4, that is, $\varepsilon \sim 0.9N(0, \sigma_1^2) + 0.1N(0, \sigma_2^2)$. For Models 5–6, set $\sigma_2^2 = 30$, and refer to cases (a) and (b) with $\sigma_1^2 = 2$ and 4, respectively. A realization of $\mathcal{X}$ for Model 5 is illustrated in Figure 5(a).

As shown in Table 3, the Lasso, Adaptive Lasso, SCAD and MCP fail to identify most of the true predictors and have a very high FNR, while the RCT is consistently the best among all these models. It indicates that the thresholding function helps us deal with these very complicated covariance structures of predictors. RCT also outperforms STGP, especially in terms of selection accuracy.

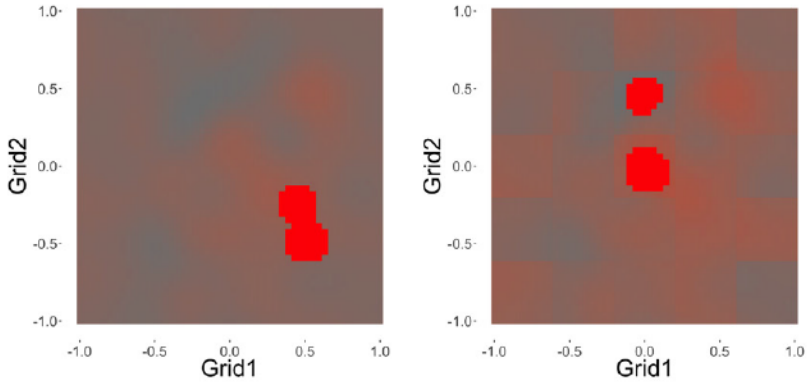
5.2. Simulation with Spatial Information

In this section, we consider the available spatial neighborhood information for the STGP and generalized RCT in Section 4. The spatial neighborhood information is useful for scalar-on-image regression applications, and the simulation results with neighborhood information can provide more practical guidance on the performance of different methods for image data analysis.

First, we consider Models 7 and 8 when there is a group structure among covariates, and the group penalty is necessary to be applied. Models 7 and 8 partition the whole image space into 25 sub-regions with equal numbers of predictors in each. The predictors from each sub-region are generated from a Gaussian process with a constant mean function and the same covariance structure as Models 5 and 6, respectively. Correlations between the magnitude of mean functions across different regions are 0.9. We randomly pick two sub-regions.

Table 3. Estimation and selection accuracy of different methods for Models 5 and 6.

	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss
	Model (5a)			Model (5b)			Model (6a)			Model (6b)		
Lasso	0.002	0.814	7.083	0.002	0.854	7.228	0.001	0.784	6.071	0.041	0.418	4.265
	(0.002)	(0.031)	(1.561)	(0.001)	(0.033)	(1.222)	(0.001)	(0.068)	(1.305)	(0.006)	(0.050)	(0.360)
AdaLasso	0.002	0.820	12.527	0.001	0.853	10.690	0.001	0.784	0.196	0.033	0.446	4.347
	(0.031)	(0.031)	(0.041)	(0.001)	(0.033)	(2.728)	(0.068)	(0.068)	(0.306)	(0.003)	(0.127)	(0.446)
MCP	0.007	0.918	7.526	0.007	0.918	7.524	0.007	0.918	7.532	0.007	0.918	7.535
	(0.003)	(0.060)	(0.622)	(0.007)	(0.060)	(0.612)	(0.003)	(0.060)	(0.655)	(0.003)	(0.069)	(0.675)
STGP	0.001	0.435	2.729	0.001	0.484	2.753	0.002	0.461	2.584	0.003	0.476	2.561
	(0.002)	(0.161)	(0.335)	(0.003)	(0.210)	(0.610)	(0.006)	(0.185)	(0.621)	(0.008)	(0.226)	(0.577)
RCT	0.025	0.018	2.302	0.034	0.016	2.761	0.027	0.196	3.038	0.045	0.303	3.284
	(0.001)	(0.041)	(0.342)	(0.008)	(0.105)	(0.298)	(0.014)	(0.306)	(1.083)	(0.016)	(0.320)	(1.215)



(a) Model 5 (b) Model 7

Figure 5. Plots of the simulated image predictors from the Gaussian process regression (Two circles of bright red color on covariates with nonzero coefficients).

Within each, we identify a circle with a randomly selected center and radius 0.13 as nonzero effect regions. This makes around 1/3 of the points within the selected sub-region have nonzero effects. We set all nonzero coefficients as 2. Mathemtically, Models 7 and 8 can be summarized as follows. For $k = 1, \ldots, 25$, let $\mathcal{X}_k(s^k) = (\mathcal{X}_k(s_1^k), \ldots, \mathcal{X}_k(s_{100}^k)) \in \mathbb{R}^{100}$ be an evaluation of Gaussian process $\mathcal{X}_k$ at 100 location points $(s_1^k, \ldots, s_{100}^k)$, which equality spaced over kth sub-region. Let $\mathcal{X} = (\mathcal{X}_1^\mathsf{T}, \ldots, \mathcal{X}_{25}^\mathsf{T})^\mathsf{T} \in (\mathbb{R}^{100})^{\otimes 25}$. The Gaussian process $\{\mathcal{X}_k(s), k = 1, \ldots, 25\}$ has the constant mean function μ_k, generated by $(\mu_1, \ldots, \mu_{25}) \sim N(0, \Gamma)$ where $\Gamma_{ij} = 0.9 + 0.1I(i = j)$, and the following covariance structures:

Model 7 (GP(10)) :

$$\kappa(s_i, s_j) = \exp(-\|s_i^k\|^2 - \|s_j^k\|^2 - 10\|s_i^k - s_j^k\|^2),$$

Model 8 (GP(5)) :

$$\kappa(s_i, s_j) = \exp(-\|s_i^k\|^2 - \|s_j^k\|^2 - 5\|s_i^k - s_j^k\|^2).$$

Figure 3(b) shows one simulated sample image predictor $\mathcal{X}$ in Model 7. For the noise term, we still set $\sigma_1^2 = 2$ and 4 as case (a) and (b), respectively, and $\sigma_2^2 = 30$.

For Models 7 and 8, as shown in Table 4, we compare performances of the Lasso, group Lasso (GLasso), sparse Group Lasso (SGL), STGP, and RCT. We include the region-level FPR (R-FPR) and region-level FNR (R-FNR) to measure the region-level selection accuracy, which are computed based on whether there is at least one variable in the region is selected. Compared with GLasso and SGL, the RCT identifies almost all the correct groups with zero false negatives and lower FPR. The RCT (info) has more precise selection accuracy by using the spatial information.

Second, we compare the performances between the STGP and RCT in the more challenging generative model structure when the spatial neighborhood information may not be helpful. We consider the following Models 9 and 10 that have a much less smooth pattern in the regression coefficients than Models 7 and 8:

Model 9 : Generate $\mathcal{X}$ and β as in Model 7, randomly keep 25% of nonzero β.

Model 10 : Generate $\mathcal{X}$ and β as in Model 8, randomly keep 25% of nonzero β.

We let the noise term follow setting (a) in Models 7 and 8. Compare with Model 7 and 8, we only choose 25% of coefficients in the selected circles to be nonzeroes. Now, the true signal does not vary smoothly across the spatial location. We summarize the results of the RCT and STGP in Table 5. We see that spatial information does not necessarily improve the numerical performance. In this scenario, RCT outperforms STGP, especially without using the spatial neighbor information.

Lastly, we demonstrate the robustness of the RCT against the STGP under the heavy-tailed error distribution. We consider the same settings of Models 5 and 6 except that under the standard Cauchy error distribution, and we denote them by Model 5 (Cauchy) and Model 6 (Cauchy). We used spatial neighborhood information for both STGP and RCT estimators. We see that when the error is heavy-tailed, the RCT significantly outperforms the STGP for both models (Table 6).

Table 4. Selection accuracy of different methods for Models 7 and 8.

	FPR	FNR	R-FPR	R-FNR	FPR	FNR	R-FPR	R-FNR
	Model (7a)				Model (8a)			
Lasso	0.019	0.270	0.101	0	0.028	0.411	0.140	0
	(0.004)	(0.072)	(0.091)	(0)	(0.006)	(0.100)	(0.083)	(0)
GLasso	0.220	0.378	0.232	0	0.215	0.403	0.232	0
	(0.055)	(0.151)	(0.094)	(0)	(0.054)	(0.141)	(0.094)	(0)
SGL	0.115	0.010	0.135	0	0.126	0.012	0.126	0
	(0.035)	(0.022)	(0.060)	(0)	(0.037)	(0.030)	(0.063)	(0)
STGP (info)	0.063	0	0.109	0	0.061	0	0.110	0
	(0.010)	(0)	(0.060)	(0)	(0.007)	(0)	(0.060)	(0)
RCT	0.059	0	0.087	0	0.064	0	0.111	0
	(0.003)	(0)	(0.087)	(0)	(0.007)	(0)	(0.091)	(0)
RCT (info)	0.051	0	0	0	0.052	0	0	0
	(0.001)	(0)	(0)	(0)	(0.001)	(0)	(0)	(0)
	Model (7b)				Model (8b)			
Lasso	0.019	0.347	0.166	0	0.028	0.415	0.145	0
	(0.004)	(0.078)	(0.088)	(0)	(0.006)	(0)	(0.096)	(0)
GLasso	0.223	0.372	0.232	0	0.214	0.405	0.232	0
	(0.053)	(0.143)	(0.094)	(0)	(0.056)	(0.140)	(0.094)	(0)
SGL	0.130	0.012	0.170	0	0.128	0.020	0.165	0
	(0.032)	(0.026)	(0.060)	(0)	(0.037)	(0.046)	(0.091)	(0)
STGP (info)	0.061	0	0.114	0	0.062	0	0.113	0
	(0.007)	(0)	(0.062)	(0)	(0.007)	(0)	(0.039)	(0)
RCT	0.066	0	0.161	0	0.065	0	0.120	0
	(0.006)	(0)	(0.100)	(0)	(0.006)	(0)	(0.096)	(0)
RCT (info)	0.051	0	0	0	0.052	0	0	0
	(0.001)	(0)	(0)	(0)	(0.001)	(0)	(0)	(0)

Table 5. Selection accuracy of STGP and RCT for Models 9 and 10.

	FPR	FNR	R-FPR	R-FNR	FPR	FNR	R-FPR	R-FNR
	Model 9				Model 10			
STGP	0.154	0.126	0.286	0.336	0.175	0.117	0.344	0.297
(info)	(0.113)	(0.106)	(0.250)	(0.206)	(0.133)	(0.092)	(0.234)	(0.179)
STGP	0.055	0.245	0.181	0.220	0.048	0.254	0.157	0.250
	(0.035)	(0.175)	(0.110)	(0.250)	(0.017)	(0.145)	(0.076)	(0.253)
RCT	0.037	0.128	0	0.245	0.036	0.139	0	0.260
(info)	(0.001)	(0)	(0)	(0)	(0.07012)	(0.078)	(0)	(0.252)
RCT	0.038	0.117	0	0.240	0.038	0.122	0	0.240
	(0.012)	(0.080)	(0)	(0.252)	(0.012)	(0.072)	(0)	(0.252)

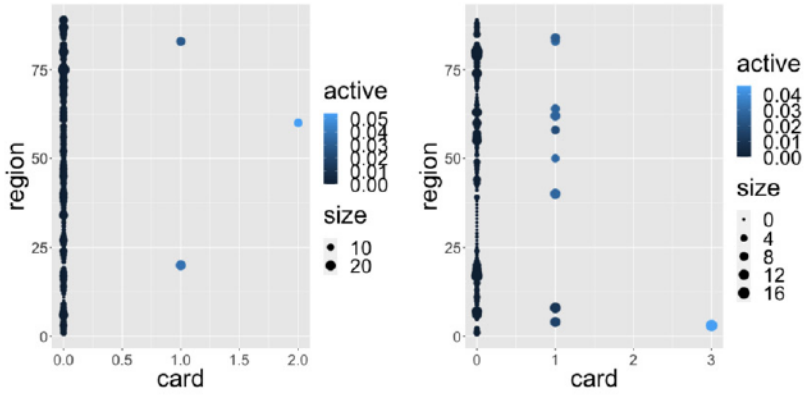
Table 6. Selection accuracy of STGP and RCT for Models 5 and 6 with the Cacuchy error.

	FPR	FNR	ℓ_2 loss	FPR	FNR	ℓ_2 loss
	Model 5 (Cauchy)			Model 6 (Cauchy)		
STGP (info)	0.144	0.300	3.075	0.131	0.386	3.369
	(0.197)	(0.470)	(0.389)	(0.185)	(0.489)	(0.825)
RCT (info)	0.005	0.116	2.210	0.005	0.071	2.187
	(0.007)	(0.184)	(0.403)	(0.006)	(0.085)	(0.220)

6. Application to Scalar-on-Image Regression Analysis

This section applies the proposed method to analyze the 2-back versus 0-back contrast maps derived from the *n*-back task fMRI imaging data in the Adolescent Brain Cognitive Development (ABCD) study (Casey et al. 2018). Our goal is to identify the important imaging features from the contrast maps that are strongly associated with the risk of psychiatric disorder, measured by the general factor of psychopathology (GFP) or "p-factor." After the standard fMRI prepossessing steps, all the images are registered into the 2 mm standard Montreal Neurological Institute (MNI) space consisting of 160,990 voxels in the 90 Automated Anatomical Labeling (AAL) brain regions. With the missing values being removed, the data used in our analysis consists of 2,070 subjects. To reduce the dimension of the imaging data, we partition 90 AAL regions into 2,518 subregions with each region consisting of an average of 64 voxels. We refer to each subregion as a super-voxel. For each subject, we compute the average intensity values of the voxels within each super-voxel as its intensity. We consider those 2518 super-voxel-wise intensity values as the potential image predictors.

There are several challenging issues in the scalar-on-image regression analysis of this dataset. First, the correlations between super-voxels across the 90 AAL regions can be very high and the correlation patterns are complex. In fact, there are 151,724 voxel pairs across these regions having a correlation larger than 0.8 (or less than −0.8), and 9,038 voxel pairs with a correlation larger than 0.9 (or less than −0.9). Figure 1 visualizes the region-wise correlation structures, where panel (a) shows the highest correlations between regions; and panel (b) counts the voxel pairs that have a correlation higher than 0.8 (or less than −0.8) in each corresponding region pair. Given the image predictors having such high and complicated covariance structures, the classical Lasso or the group Lasso method may fail to perform variable selection satisfactorily. In contrast, the proposed model with coefficient thresholding is developed to resolve this issue since it does not require the strong conditions on the design matrix. Second, the AAL brain atlas provides useful information on the brain structure and function that may be related to the risk of psychiatric disorders. It is of interest to integrate the AAL region partition as grouping information of image predictors to improve the accuracy of imaging feature selection. Third, the outcome variable "p-factor" has a heavy tail with a kurtosis of 66. Compared with normal distribution having a kurtosis of 3, our outcome variable is heavy-tailed with potential outliers. The existing nonrobust scalar-on-image regression methods may produce inaccurate results. All the aforementioned challenging



(a) Soft Thresholded Gaussian Process (b) Robust Coefficient Thresholding

Figure 6. Illustration of the selection frequency and cardinality of 90 brain regions for STGP and the RCT. The X axis shows the maximum selection frequency, the bubble size is proportional to the number of super-voxels with the selection frequency being larger than 0.6, and the color indicates the proportion of super-voxels for each brain region.

issues motivate the needs of developing our robust regression with coefficient thresholding and group penalty.

In our analysis, we adjust confounding effects by including a few predictors in the model: family size, gender, race, highest parents' education, household marital status and household income level. Given the intrinsic group structure, we compare the performance of the proposed method and the STGP in this data analysis. As we mentioned, the fMRI data analysis generally suffers from the reliability issue due to its complex data structure and low signal-to-noise ratio (Bennett and Miller 2010; Brown and Behrmann 2017; Eklund et al. 2012; Eklund, Nichols, and Knutsson 2016). To evaluate the variable selection stability for both methods, we consider a bootstrap approach with 100 replications. In each replication, we sample n observations with replacement, and fit the bootstrap samples using the best set of tuning parameters chosen by a 5-fold cross-validation. Then we obtain the frequency of each super-voxel being selected over 100 replications as a measure of the selection stability, which can be used to fairly compare the regions that can be consistently selected against randomness, and thus ensure the reliability of the scientific findings in our analysis.

RCT and STGP, respectively, select 124.5 and 245.3 super-voxels per replication on average. Figure 6 displays the bootstrap selection results, where the x-axis represents the maximum selection frequency of super-voxels in each region. The circle size is proportional to the number of super-voxels with the corresponding selection frequency being larger than 0.6. The color represents the proportion of super-voxels being selected in each region. Despite a smaller number of super-voxels being selected in each bootstrap run, RCT consistently selects super-voxels in several important brain regions over bootstrap samples, while STGP identifies a less number of brain regions that contain selected super-voxels.

Table 7 summarizes the comparisons of selected regions from RCT and STGP by varying different thresholds of selection frequency from 0.6 to 0.9. Compared with STGP, RCT selects more stable regions for each level of selection frequency, indicating that our method produces more reliable selection results. In particular, containing at least one super-voxels with more than 60% selection frequency, seven and three regions are respectively, identified by RCT and STGP. Among those regions,

Table 7. Comparisons of stable selection regions between RCT and STGP for different levels of selection frequency threshold.

Selected frequency	RCT	STGP
0.6–0.7	Callcarine_L, Occipital_Mid_L, Parietal_Inf_L	Frontal_Sup_Midial_R, Temporal_Mid_L
0.7–0.8	Temporal_Mid_L	SupraMarginal_R
0.8–0.9	Frontal_Mid_Orb_L, Precuneus_L	N/A
>0.9	Frontal_Sup_L	N/A

only one common region, that is, the left temporal gyrus (Temporal_Mid_L), is detected by both methods, where RCT has a higher selection frequency (0.79) than STGP (0.69). The existing functional neuroimaging studies have indicated that the middle temporal gyrus is involved in language and semantic memory processing (Cabeza and Nyberg 2000), and it is also related to mental diseases such as chronic schizophrenia (Onitsuka et al. 2004).

Among other selected regions, with more than 90% selection frequency, RCT consistently selects super-voxels in the left superior frontal gyrus (Frontal_Sup_L), while the selection frequency by STGP is below 60%. Superior frontal gyrus is known to be strongly related to working memory (Boisgueheneuc et al. 2006) which plays a critical role in attending to and analyzing incoming information. Deficits in working memory are associated with many cognitive and mental health challenges, such as anxiety and stress (Lukasik et al. 2019), which can be captured by the "p-factor." The strong relationship between p-factor and working memory has been discovered by existing studies (Huang-Pollock et al. 2017).

In addition, RCT also identifies five more regions than STGP: precuneus, the left middle frontal gyrus (Frontal_Mid_Orb_L), the left alcarine fissure and surrounding cortex (Calcarine_L), the middle occipital gyrus (Occipital_Mid_L) and the left inferior parietal gyri (Parietal_Inf_L). Percuneus is well studied as a core of mind (Cavanna and Trimble 2006), and it is highly related to posttraumatic stress disorder(PTSD) and other mental health issue (Geuze et al. 2007). Middle frontal gyrus is part of limbic system and known to be highly related to emotion (Sprooten et al. 2017). Inferior parietal lobule has been involved in the perception of emotions in facial stimuli,

and interpretation of sensory information (Radua et al. 2010). Calcarine fissure is related to vision. Middle occipital gyrus is primarily responsible for object recognition. Our findings are supported by the existing study on brain AAL regions such as Power et al. (2011). Specifically, among the detected regions, Frontal_Sup_L, Frontal_Mid_Orb_L and Parietal_Inf_L are related to the task control network, and Calcarine_L and Ociptial_Mid_L are both in the visual network. This makes sense as the working memory task is the visual task. We also investigate the empirical correlation between the selected region and p-factor. For the most frequently selected regions Frontal_Sup_L, Frontal_Mid_Orb_L and Parietal_Inf_L which are related to the task control network, 28 out of 108 super voxels are significantly correlated with p-factor at the significant level of 0.1 based on the Kendall's rank correlation test. It would be interesting to further investigate how the brain activity in these regions influences the p-factor.

To further demonstrate the proposed method providing more reliable scientific findings in comparison to STGP, we evaluate the prediction performance of the two methods. We randomly split the data into two parts with 80% as the training data for model fitting and 20% as the test data for computing the prediction error. We repeat this procedure for 50 times. The mean absolute prediction error of the RCT is 0.464 with standard error 0.004, while the STGP has a mean absolute prediction error of 0.480 with standard error 0.038. Compared with the STGP, our proposed method improves the prediction performance of the p-factor using working memory contrast maps in the ABCD study.

7. Conclusion

In this article, we propose a novel high-dimensional robust regression with coefficient thresholding in the presence of complex dependencies among predictors and potential outliers. The proposed method uses the power of thresholding functions and the robust Huber loss to build an efficient nonconvex estimation procedure. We carefully analyze the landscape of the nonconvex loss function for the proposed method, which enables us to establish both statistical and computational consistency. We also present an extension to incorporate the spatial information into the proposed method. We demonstrate the effectiveness and usefulness of the proposed method in simulation studies and a real application to imaging data analysis. In the future, it is interesting to investigate how to incorporate the spatial-temporal information of the imaging data into our proposed method. It is also important to study the statistical consistency of the near-stationary solution from the proposed gradient descent based algorithm under more general conditions.

Supplementary Materials

The supplementary materials provide technical remarks, the proofs of lemmas and theorems, and additional numerical results.

Acknowledgments

The authors would like to thank the editor, associate editor, and three referees for their insightful comments and constructive suggestions that improved the quality and presentation of this article.

Funding

Bingyuan Liu, Qi Zhang, and Lingzhou Xue were partially supported by the National Science Foundation (NSF) grants DMS-1811552, DMS-1953189, DMS-2210775, and an National Institutes of Health (NIH) grant R21AI144765. Song's research was partially supported by an NSF grant DMS-1811734. Kang's research was partially supported by an NSF grant IIS-2123777 and the NIH grants R01DA048993, R01MH105561, and R01GM124061.

ORCID

Lingzhou Xue http://orcid.org/0000-0002-8252-0637
Jian Kang http://orcid.org/0000-0002-5643-2668

References

Barron, A., Birgé, L., and Massart, P. (1999), "Risk Bounds for Model Selection via Penalization," *Probability Theory and Related Fields*, 113, 301–413. [715,717]

Bennett, C. M., and Miller, M. B. (2010), "How Reliable are the Results from Functional Magnetic Resonance Imaging?" *Annals of the New York Academy of Sciences*, 1191, 133–155. [716,727]

Boisgueheneuc, F. d., Levy, R., Volle, E., Seassau, M., Duffau, H., Kinkingnehun, S., Samson, Y., Zhang, S., and Dubois, B. (2006), "Functions of the Left Superior Frontal Gyrus in Humans: A Lesion Study," *Brain*, 129, 3315–3328. [727]

Brown, E. N., and Behrmann, M. (2017), "Controversy in Statistical Analysis of Functional Magnetic Resonance Imaging Data," *Proceedings of the National Academy of Sciences*, 114, E3368–E3369. [716,727]

Cabeza, R., and Nyberg, L. (2000), "Imaging Cognition II: An Empirical Review of 275 Pet and fMRI Studies," *Journal of Cognitive Neuroscience*, 12, 1–47. [727]

Cai, T. T., Li, X., and Ma, Z. (2016), "Optimal Rates of Convergence for Noisy Sparse Phase Retrieval via Thresholded Wirtinger Flow," *The Annals of Statistics*, 44, 2221–2251. [717]

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015), "Phase Retrieval via Wirtinger Flow: Theory and Algorithms," *IEEE Transactions on Information Theory*, 61, 1985–2007. [717]

Casey, B., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H. et al. (2018), "The Adolescent Brain Cognitive Development (abcd) Study: Imaging Acquisition across 21 Sites," *Developmental Cognitive Neuroscience*, 32, 43–54. [715,726]

Cavanna, A. E., and Trimble, M. R. (2006), "The Precuneus: A Review of its Functional Anatomy and Behavioural Correlates," *Brain*, 129, 564–583. [727]

Charbonnier, P., Blanc-Féraud, L., Aubert, G., and Barlaud, M. (1997), "Deterministic Edge-Preserving Regularization in Computed Imaging," *IEEE Transactions on Image Processing*, 6, 298–311. [716,718]

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., and Knutsson, H. (2012), "Does Parametric fMRI Analysis with SPM Yield Valid Results? – An Empirical Study of 1484 Rest Datasets," *NeuroImage*, 61, 565–578. [716,727]

Eklund, A., Nichols, T. E., and Knutsson, H. (2016), "Cluster Failure: Why fMRI Inferences for Spatial Extent have Inflated False-Positive Rates," *Proceedings of the National Academy of Sciences*, 113, 7900–7905. [716,727]

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013), "On Robust Regression with High-Dimensional Predictors," *Proceedings of the National Academy of Sciences*, 110, 14557–14562. [718]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [715]

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018), "I-lamm for Sparse Learning: Simultaneous Control of Algorithmic Complexity and Statistical Error," *The Annals of Statistics*, 46, 814–841. [716]

Fan, J., Xue, L., and Zou, H. (2014), "Strong Oracle Optimality of Folded Concave Penalized Estimation," *The Annals of Statistics*, 42, 819–849. [715,716,724]

Fan, Y., and Lv, J. (2013), "Asymptotic Equivalence of Regularization Methods in Thresholded Parameter Space," *Journal of the American Statistical Association*, 108, 1044–1061. [716,717]

Geuze, E., Vermetten, E., de Kloet, C. S., and Westenberg, H. G. (2007), "Precuneal Activity during Encoding in Veterans with Posttraumatic Stress Disorder," *Progress in Brain Research*, 167, 293–297. [727]

Ghadimi, S., Lan, G., and Zhang, H. (2016), "Mini-Batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization," *Mathematical Programming*, 155, 267–305. [722]

Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2014), "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection," *Journal of Computational and Graphical Statistics*, 23, 46–64. [716]

He, K., Xu, H., and Kang, J. (2018), "A Selective Overview of Feature Screening Methods with Applications to Neuroimaging Data," *Wiley Interdisciplinary Reviews: Computational Statistics*, 11, e1454. [715,719,724]

Hochreiter, S. (1998), "The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6, 107–116. [722]

Huang-Pollock, C., Shapiro, Z., Galloway-Long, H., and Weigard, A. (2017), "Is Poor Working Memory a Transdiagnostic Risk Factor for Psychopathology?," *Journal of Abnormal Child Psychology*, 45, 1477–1490. [727]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 53, 73–101. [718]

Jain, P., Tewari, A., and Kar, P. (2014), "On Iterative Hard Thresholding Methods for High-Dimensional m-estimation," in *Advances in Neural Information Processing Systems* (Vol. 27). [716,717]

Kang, J., Reich, B. J., and Staicu, A.-M. (2018), "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process," *Biometrika*, 105, 165–184. [715,716,717,719,722]

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. (2015), "Spatial Bayesian Variable Selection and Grouping for High-Dimensional Scalar-on-Image Regression," *The Annals of Applied Statistics*, 9, 687–713. [716]

Lindquist, M. A. (2008), "The Statistical Analysis of fMRI Data," *Statistical Science*, 23, 439–464. [716]

Loh, P.-L. (2017), "Statistical Consistency and Asymptotic Normality for High-Dimensional Robust m-estimators," *The Annals of Statistics*, 45, 866–896. [718]

——— (2018), "Scale Calibration for High-Dimensional Robust Regression," arXiv preprint arXiv:1811.02096. [718]

Loh, P.-L., and Wainwright, M. J. (2013), "Regularized m-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima," in *Advances in Neural Information Processing Systems*, pp. 476–484. [717]

——— (2017), "Support Recovery without Incoherence: A Case for Nonconvex Regularization," *The Annals of Statistics*, 45, 2455–2482. [717,718,724]

Lukasik, K. M., Waris, O., Soveri, A., Lehtonen, M., and Laine, M. (2019), "The Relationship of Anxiety and Stress with Working Memory Performance in a Large Non-Depressed Sample," *Frontiers in Psychology*, 10, 4. [727]

Mei, S., Bai, Y., and Montanari, A. (2018), "The Landscape of Empirical Risk for Nonconvex Losses," *The Annals of Statistics*, 46, 2747–2774. [717,718,721]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of m-estimators with Decomposable Regularizers," *Statistical Science*, 27, 538–557. [717]

Negahban, S., and Wainwright, M. J. (2012), "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise," *Journal of Machine Learning Research*, 13, 1665–1697. [717]

Nesterov, Y. (2013), "Gradient Methods for Minimizing Composite Functions," *Mathematical Programming*, 140, 125–161. [721,723]

Onitsuka, T., Shenton, M. E., Salisbury, D. F., Dickey, C. C., Kasai, K., Toner, S. K., Frumin, M., Kikinis, R., Jolesz, F. A., and McCarley, R. W. (2004), "Middle and Inferior Temporal Gyrus Gray Matter Volume Abnormalities in Chronic Schizophrenia: An MRI Study," *American Journal of Psychiatry*, 161, 1603–1611. [727]

Poldrack, R. A. (2012), "The Future of fMRI in Cognitive Neuroscience," *Neuroimage*, 62, 1216–1220. [716]

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., and Petersen, S. E. (2011), "Functional Network Organization of the Human Brain," *Neuron*, 72, 665–678. [728]

Radua, J., Phillips, M. L., Russell, T., Lawrence, N., Marshall, N., Kalidindi, S., El-Hage, W., McDonald, C., Giampietro, V., Brammer, M. J., David, A. S., Surguladze, S. A. (2010), "Neural Response to Specific Components of Fearful Faces in Healthy and Schizophrenic Adults," *Neuroimage*, 49, 939–946. [728]

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., and Chun, M. M. (2016), "A Neuromarker of Sustained Attention from Whole-Brain Functional Connectivity," *Nature Neuroscience*, 19, 165–171. [716]

Shi, R., and Kang, J. (2015), "Thresholded Multiscale Gaussian Processes with Application to Bayesian Feature Selection for Massive Neuroimaging Data," arXiv preprint arXiv:1504.06074 . [716,717]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [718]

Sprooten, E., Rasgon, A., Goodman, M., Carlin, A., Leibu, E., Lee, W. H., and Frangou, S. (2017), "Addressing Reverse Inference in Psychiatric Neuroimaging: Meta-Analyses of Task-Related Brain Activation in Common Mental Disorders," *Human Brain Mapping*, 38, 1846–1864. [727]

Sun, Q., Jiang, B., Zhu, H., and Ibrahim, J. G. (2019), "Hard Thresholding Regression," *Scandinavian Journal of Statistics*, 46, 314–328. [716,717]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [715]

Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery using ℓ_{1}-constrained Quadratic Programming (lasso)," *IEEE Transactions on Information Theory*, 55, 2183–2202. [715]

Wang, L., Chen, G., and Li, H. (2007), "Group Scad Regression Analysis for Microarray Time Course Gene Expression Data," *Bioinformatics*, 23, 1486–1494. [718]

Wang, X., Zhu, H., and Initiative, A. D. N. (2017), "Generalized Scalar-on-Image Regression Models via Total Variation," *Journal of the American Statistical Association*, 112, 1156–1168. [715,716]

Yang, Y. (2015), "A Unified Algorithm for Fitting Penalized Models with High Dimensional Data," PhD thesis, University of Minnesota. [723]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [718]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [715,724]

Zhang, C.-H., and Zhang, T. (2012), "A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems," *Statistical Science*, 27, 576–593. [717]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [715]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [715,718]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533. [715,716]