THE LASSO FOR HIGH-DIMENSIONAL REGRESSION WITH A POSSIBLE CHANGE-POINT

SOKBAE LEE, MYUNG HWAN SEO, AND YOUNGKI SHIN

Abstract. We consider a high-dimensional regression model with a possible change-point due to a covariate threshold and develop the Lasso estimator of regression coefficients as well as the threshold parameter. Under a sparsity assumption, we derive nonasymptotic oracle inequalities for both the prediction risk and the ℓ_1 estimation loss for regression coefficients. Since the Lasso estimator selects variables simultaneously, we show that oracle inequalities can be established without pretesting the existence of the threshold effect. Therefore, the Lasso estimator not only selects covariates but also accomplishes model selection between the linear and threshold regression models. Furthermore, we establish conditions under which the unknown threshold parameter can be estimated at a rate of nearly n^{-1} when the number of regressors can be much larger than the sample size (n). We illustrate the usefulness of our proposed estimation method via Monte Carlo simulations and an application to real data.

Key words. Lasso, oracle inequalities, sample splitting, sparsity, threshold models.

AMS Subject Classification. Primary 62H12, 62J05; secondary 62J07.

1. Introduction

The Lasso and related methods have received rapidly increasing attention in statistics since the seminal work of Tibshirani (1996). For example, see a timely monograph by Bühlmann and van de Geer (2011) as well as review articles by Fan and Lv (2010) and Tibshirani (2011) for general overview and recent developments.

In this paper, we develop a method for estimating a high-dimensional regression model with a possible change-point due to a covariate threshold, while selecting relevant regressors from a set of many potential covariates. In particular, we propose the ℓ_1 penalized least squares (Lasso) estimator of parameters, including the unknown

Date: 31 December 2012.
Supported in part by the European Research Council (ERC-2009-StG-240910-ROMETA).

threshold parameter, and analyze its properties under a sparsity assumption when the number of possible covariates can be much larger than the sample size.

To be specific, let $\{(Y_i, X_i, Q_i) : i = 1, \ldots, n\}$ be a sample of independent observations such that

$$Y_i = X_i'\beta_0 + X_i'\delta_0 1\{Q_i < \tau_0\} + U_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where for each i, X_i is an $M \times 1$ deterministic vector, Q_i is a deterministic scalar, U_i follows $N(0, \sigma^2)$, and $1\{\cdot\}$ denotes the indicator function. The scalar variable Q_i is the threshold variable and τ_0 is the unknown threshold parameter. Note that since Q_i is a fixed variable in our setup, (1.1) includes a regression model with a change-point at unknown time (e.g. $Q_i = i/n$).

A regression model such as (1.1) offers applied researchers a simple yet useful framework to model nonlinear relationships by splitting the data into subsamples. Empirical examples include cross-country growth models with multiple equilibria (Durlauf and Johnson, 1995), racial segregation (Card et al., 2008), and financial contagion (Pesaran and Pick, 2007), among many others. Typically, the choice of the threshold variable is well motivated in applied work (e.g. initial per capita output in Durlauf and Johnson (1995), and the minority share in a neighborhood in Card et al. (2008)), but selection of other covariates is subject to applied researchers' discretion. However, covariate selection is important in identifying threshold effects (i.e., nonzero δ_0) since a piece of evidence favoring threshold effects with a particular set of covariates could be overturned by a linear model with a broader set of regressors. Therefore, it seems natural to consider Lasso as a tool to estimate (1.1).

The statistical problem we consider in this paper is to estimate unknown parameters $(\beta_0, \delta_0, \tau_0) \in \mathbb{R}^{2M+1}$ when M is much larger than n. For the classical setup (estimation of parameters without covariate selection when M is smaller than n), estimation of

(1.1) has been well studied (e.g. Tong, 1990; Chan, 1993; Hansen, 2000). Also, a general method for testing threshold effects in regression (i.e. testing $H_0 : \delta_0 = 0$ in (1.1)) is available for the classical setup (e.g. Lee et al., 2011).

Although there are many papers on Lasso type methods and also equally many papers on change points, sample splitting, and threshold models, there seem to be only a handful of papers that intersect both topics. Wu (2008) proposed an information-based criterion for carrying out change point analysis and variable selection simultaneously in linear models with a possible change point; however, the proposed method in Wu (2008) would be infeasible in a sparse high-dimensional model. In change-point models without covariates, Harchaoui and Levy-Leduc (2008, 2010) proposed a method for estimating the location of change-points in one-dimensional piecewise constant signals observed in white noise, using a penalized least-square criterion with an ℓ_1-type penalty, and Zhang and Siegmund (2007) developed Bayes Information Criterion (BIC)-like criteria for determining the number of changes in the mean of multiple sequences of independent normal observations when the number of change-points can increase with the sample size. Ciuperca (2012) considered a similar estimation problem as ours, but the corresponding analysis is restricted to the case when the number of potential covariates is small.

In this paper, we consider the Lasso estimator of regression coefficients as well as the threshold parameter. Since the change-point parameter τ_0 does not enter additively in (1.1), the resulting optimization problem in the Lasso estimation is non-convex. We overcome this problem by comparing the values of standard Lasso objective functions on a grid over the range of possible values of τ_0.

Theoretical properties of the Lasso and related methods for high-dimensional data are examined by Fan and Peng (2004), Bunea et al. (2007), Candès and Tao (2007), Huang et al. (2008), Huang et al. (2008), Kim et al. (2008), Bickel et al. (2009), and Meinshausen and Yu (2009), among many others. Most of the papers consider quadratic objective functions and linear or nonparametric models with an additive mean zero error. There has been recent interest in extending this framework to generalized linear models (e.g. van de Geer, 2008; Fan and Lv, 2011), to quantile regression models (e.g. Belloni and Chernozhukov, 2011a; Bradic et al., 2011; Wang et al., 2012), and to hazards models (e.g. Bradic et al., 2012; Lin and Lv, 2012). We contribute to this literature by considering a regression model with a possible change-point and then deriving nonasymptotic oracle inequalities for both the prediction risk and the ℓ_1 estimation loss for regression coefficients under a sparsity scenario.

Our theoretical results build on Bickel et al. (2009). Since the Lasso estimator selects variables simultaneously, we show that oracle inequalities similar to those obtained in Bickel et al. (2009) can be established without pretesting the existence of the threshold effect. In particular, when there is no threshold effect ($\delta_0 = 0$), we prove oracle inequalities that are basically equivalent to those in Bickel et al. (2009). Therefore, the Lasso estimator not only selects covariates but also accomplishes model selection between the linear and threshold regression models. Furthermore, when $\delta_0 \neq 0$, we establish conditions under which the unknown threshold parameter can be estimated at a rate of nearly n^{-1} when the number of regressors can be much larger than the sample size. To achieve this, we develop some sophisticated chaining arguments and provide sufficient regularity conditions under which we prove oracle inequalities. The super-consistency result of $\widehat{\tau}$ is well known when the number of covariates is small (see, e.g. Chan, 1993). To the best of our knowledge, our paper is the first work that demonstrates the possibility of a nearly n^{-1} bound in the context of sparse high-dimensional regression models with a change-point.

The remainder of this paper is as follows. In Section 2 we propose the Lasso estimator, and in Section 3 we give a brief illustration of our proposed estimation method using a real-data example in economics. In Section 4 we establish the prediction consistency of our Lasso estimator. In Section 5, we establish sparsity oracle inequalities in terms of both the prediction loss and the ℓ_1 estimation loss for (α_0, τ_0), while providing low-level sufficient conditions for two possible cases of threshold effects. In Section 6 we present results of some simulation studies, and Section 7 concludes. Appendix A gives some additional discussions on identifiability for τ_0 and Appendices B and C contain all the proofs. Throughout the paper, let $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$ for any real numbers a and b.

2. Lasso Estimation

Let $\mathbf{X}_i(\tau)$ denote the $(2M \times 1)$ vector such that $\mathbf{X}_i(\tau) = (X_i', X_i' 1\{Q_i < \tau\})'$ and let $\mathbf{X}(\tau)$ denote the $(n \times 2M)$ matrix whose i-th row is $\mathbf{X}_i(\tau)'$. Let $\alpha_0 = (\beta_0', \delta_0')'$. Then (1.1) can be written as

$$Y_i = \mathbf{X}_i(\tau_0)'\alpha_0 + U_i, \quad i = 1, \ldots, n. \tag{2.1}$$

Following Bickel et al. (2009), we use the following notation. For an L-dimensional vector a, let $|a|_p$ denote the ℓ_p norm of a, and $|J|$ denote the cardinality of J, where $J(a) = \{j \in \{1, \ldots, L\} : a_j \neq 0\}$. In addition, let $\mathcal{M}(a)$ denote the number of nonzero elements of a. Then,

$$\mathcal{M}(a) = \sum_{j=1}^{L} 1\{a_j \neq 0\} = |J(a)|.$$

The value $\mathcal{M}(\alpha_0)$ characterizes the sparsity of the model (2.5). Also, let a_J denote the vector in $\mathbb{R}^L$ that has the same coordinates as a on J and zero coordinates on the complement J^c of J. For any n-dimensional vector $W = (W_1, \ldots, W_n)'$, define the empirical norm as

$$\|W\|_n := \left(n^{-1} \sum_{i=1}^{n} W_i^2 \right)^{1/2}.$$

Let $\mathbf{y} \equiv (Y_1, \ldots, Y_n)'$. For any fixed τ, consider the residual sum of squares

$$\begin{aligned} S_n(\alpha,\tau) &= n^{-1}\sum_{i=1}^{n}\left(Y_i - X_i'\beta - X_i'\delta 1\{Q_i < \tau\}\right)^2 \\ &= \left\|\mathbf{y} - \mathbf{X}(\tau)\alpha\right\|_n^2, \end{aligned}$$

where $\alpha = (\beta', \delta')'$.

Indicating by the superscript $^{(j)}$ the j-th element of a vector or the j-th column of a matrix, define the following $(2M \times 2M)$ diagonal matrix:

$$\mathbf{D}(\tau) := \operatorname{diag}\left\{\left\|\mathbf{X}(\tau)^{(j)}\right\|_n, \quad j = 1, ..., 2M\right\}.$$

For each fixed τ, define the Lasso solution $\widehat{\alpha}(\tau)$ by

$$\widehat{\alpha}(\tau) := \operatorname{argmin}_{\alpha \in \mathbb{R}^{2M}} \left\{S_n(\alpha,\tau) + \lambda \left|\mathbf{D}(\tau)\alpha\right|_1\right\}, \tag{2.2}$$

where λ is a tuning parameter that depends on n. It is important to note that for each fixed τ, $\widehat{\alpha}(\tau)$ is the weighted Lasso, which has advantages over the unweighted Lasso since different values of τ generate different dictionaries.

We now estimate τ_0 by

$$\widehat{\tau} := \operatorname{argmin}_{\tau \in \mathbb{T} \subset \mathbb{R}} \left\{S_n(\widehat{\alpha}(\tau),\tau) + \lambda \left|\mathbf{D}(\tau)\widehat{\alpha}(\tau)\right|_1\right\}, \tag{2.3}$$

where $\mathbb{T} \equiv [t_0, t_1]$ is a parameter space for τ_0. In fact, for any finite n, $\widehat{\tau}$ is given by an interval and we simply define the maximum of the interval as our estimator. If we wrote the model using $1\{Q_i > \tau\}$, then the convention would be the minimum of the interval being the estimator. Then the estimator of α_0 is defined as $\widehat{\alpha} := \widehat{\alpha}(\widehat{\tau})$. In fact, our proposed estimator of (α, τ) can be viewed as the one-step minimizer such that:

$$(\widehat{\alpha}, \widehat{\tau}) := \operatorname{argmin}_{\alpha \in \mathbb{R}^{2M}, \tau \in \mathbb{T} \subset \mathbb{R}} \left\{ S_n(\alpha, \tau) + \lambda \left| \mathbf{D}(\tau)\alpha \right|_1 \right\}. \tag{2.4}$$

It is worth noting that we penalize β_0 and δ_0 in (2.4), where δ_0 is the change of regression coefficients between two regimes. The model in (1.1) can be written as

$$\begin{aligned} Y_i &= X_i'\beta_0 + U_i, \quad \text{if } Q_i \geq \tau_0, \\ Y_i &= X_i'\beta_1 + U_i, \quad \text{if } Q_i < \tau_0, \end{aligned} \tag{2.5}$$

where $\beta_1 \equiv \beta_0 + \delta_0$. In view of (2.5), alternatively, one might penalize β_0 and β_1 instead of β_0 and δ_0. We opted to penalize δ_0 in this paper since and the $\delta_0 = 0$ case corresponds to the linear model.

3. Empirical Illustration

In this section, we apply the proposed Lasso method to growth regression models in economics. The neoclassical growth model predicts that economic growth rates converge in the long run. This theory has been tested empirically by looking at the negative relationship between the long-run growth rate and the initial GDP given other covariates (see Barro and Sala-i-Martin (1995) and Durlauf et al. (2005) for literature reviews). Although empirical results confirmed the negative relationship between the growth rate and the initial GDP, there has been some criticism that the results heavily depend on the selection of covariates. Recently, Belloni and Chernozhukov (2011b) show that the Lasso estimation can help select the covariates in the *linear* growth regression model and that the Lasso estimation results reconfirm the negative relationship between the long-run growth rate and the initial GDP.

We consider the growth regression model with a possible threshold. Durlauf and Johnson (1995) provide the theoretical background of the existence of multiple steady states and estimate the model with two possible threshold variables. They check the robustness by adding other available covariates in the model, but it is not still free from the criticism of the *ad hoc* variable selection. Our proposed Lasso method might be a good alternative in this situation. Furthermore, as we will show later, our method works well even if there is no threshold effect in the model. Therefore, one might expect more robust results from our approach.

The regression model we consider has the following form:

$$gr_i = \beta_0 + \beta_1 lgdp60_i + X_i'\beta_2 + 1\{Q_i < \tau\}\left(\delta_0 + \delta_1 lgdp60_i + X_i'\delta_2\right) + \varepsilon_i \tag{3.1}$$

where gr_i is the annualized GDP growth rate of country i from 1960 to 1985, $lgdp60_i$ is the log GDP in 1960, and Q_i is a possible threshold variable for which we use the initial GDP and the adult literacy rate in 1960 following Durlauf and Johnson (1995). Finally, X_i is a vector of additional covariates related to education, market efficiency, political stability, market openness, and demographic characteristics. Table 1 gives the list of all covariates used and the description of each variable. We include as many covariates as possible, which might mitigate the potential omitted variable bias. The data set mostly comes from Barro and Lee (1994), and the additional adult literacy rate is from Durlauf and Johnson (1995). Because of missing observations, we have 80 observations with 46 covariates (including a constant term) when Q_i is the initial GDP ($n = 80$ and $M = 46$), and 70 observations with 47 covariates when Q_i is the literacy rate ($n = 70$ and $M = 47$).

Tables 2 and 3 summarize the model selection and estimation results. To compare different model specifications, we also apply the Lasso procedure to a linear model, i.e. all δ's are zeros in Equation (3.1). In each case, the regularization parameter λ is chosen by the 'leave-one-out' least squares cross validation method.

Main empirical findings are as follows. First, note that the number of covariates in the threshold models is bigger than the number of observations ($2M > n$ in our notation). Thus, we cannot adopt the standard least squares method to estimate the threshold regression model. Second, the coefficients of *lgdp60* are negative in all models, which confirms the theory of the neoclassical growth model. Third, the coefficients of interaction terms between *lgdp60* and various education variables show the existence of threshold effects in both threshold model specifications. This result implies that the growth convergence rates can vary according to different education levels. Specifically, note that the interaction term between *lgdp60* and '*educ*' implies that the marginal effect of *lgdp60* becomes

$$\frac{\partial gr}{\partial lgdp60} = \beta_1 + \beta_2 educ + 1\{Q < \gamma\}(\delta_1 + \delta_2 educ).$$

In both threshold models, we have $\delta_1 = 0$, but some δ_2's are not zero. Thus, conditional on other covariates, there exist different technological diffusion effects according to the threshold point. In other words, a country with high education levels will converge faster by absorbing technology easily and quickly. Finally, the Lasso with the threshold model specification selects a more parsimonious model than that with the linear specification even though the former imposes more covariates.

Compared to the results by Durlauf and Johnson (1995), our estimation results show a couple of different points. The Lasso estimator does not confirm the threshold effect for the variable *lgdp60* itself. Different convergent rates are made only through the interaction with the education variables. It is also noteworthy that the threshold parameter estimates are much higher than those chosen by Durlauf and Johnson (1995). These differences show the importance of model selection and the advantage of the proposed Lasso estimation.

4. Prediction Consistency

In this section, we establish the prediction consistency of our Lasso estimator. For notational simplicity, we make the following convention, that is, $\widehat{\mathbf{D}} = \mathbf{D}(\widehat{\tau})$ and $\mathbf{D} = \mathbf{D}(\tau_0)$, and similarly, $\widehat{S}_n = S_n(\widehat{\alpha}, \widehat{\tau})$ and $S_n = S_n(\alpha_0, \gamma_0)$, and so on.

Define $f_{(\alpha,\tau)}(x,q) := x'\beta + x'\delta 1\{q < \tau\}$, $f_0(x,q) := x'\beta_0 + x'\delta_0 1\{q < \tau_0\}$, and $\widehat{f}(x,q) := x'\widehat{\beta} + x'\widehat{\delta}1\{q < \widehat{\tau}\}$. Let

$$V_{1j} := \left(n\sigma \left\|X^{(j)}\right\|_n\right)^{-1} \sum_{i=1}^n U_i X_i^{(j)},$$

$$V_{2j}(\tau) := \left(n\sigma \left\|X^{(j)}(\tau)\right\|_n\right)^{-1} \sum_{i=1}^n U_i X_i^{(j)} 1\{Q_i < \tau\}.$$

For a constant $\mu \in (0,1)$, define the events

$$\mathbb{A} := \bigcap_{j=1}^M \{2|V_{1j}| \le \mu\lambda/\sigma\},$$

$$\mathbb{B} := \bigcap_{j=1}^M \left\{2 \sup_{\tau\in\mathbb{T}} |V_{2j}(\tau)| \le \mu\lambda/\sigma\right\},$$

Also define $J_0 := J(\alpha_0)$ and $R_n := R_n(\alpha_0, \tau_0)$, where

$$R_n(\alpha,\tau) := 2n^{-1}\sum_{i=1}^n U_i X_i'\delta\left\{1(Q_i < \widehat{\tau}) - 1(Q_i < \tau)\right\}.$$

The following lemma gives some useful inequalities that provide a basis for all our theoretical results.

Lemma 1 (Basic Inequalities). *Conditional on the events $\mathbb{A}$ and $\mathbb{B}$, we have*

$$\left\|\widehat{f} - f_0\right\|_n^2 + (1-\mu)\lambda\left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)\right|_1 \le 2\lambda\left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0}\right|_1 + \lambda\left|\left|\widehat{\mathbf{D}}\alpha_0\right|_1 - \left|\mathbf{D}\alpha_0\right|_1\right| + R_n \tag{4.1}$$

and

$$\left\| \widehat{f} - f_0 \right\|_n^2 + (1-\mu)\,\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0) \right|_1 \le 2\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0} \right|_1 + \left\| f_{(\alpha_0, \widehat{\tau})} - f_0 \right\|_n^2 . \tag{4.2}$$

The basic inequalities in Lemma 1 involve more terms than that of the linear model (e.g. Lemma 6.1 of Bühlmann and van de Geer, 2011) because our model in (1.1) includes the unknown threshold parameter τ_0 and the weighted Lasso is considered in (2.2). Also, it helps prove our main results to have different upper bounds in (4.1) and (4.2) for the same lower bound.

We now establish conditions under which $\mathbb{A} \cap \mathbb{B}$ has probability close to one with a suitable choice of λ. Define

$$r_n := \min_{1 \le j \le M} \frac{\left\| X^{(j)}(t_0) \right\|_n^2}{\left\| X^{(j)} \right\|_n^2}, \tag{4.3}$$

where $X^{(j)}(\tau) \equiv (X_1^{(j)} 1\{Q_1 < \tau\}, \ldots, X_n^{(j)} 1\{Q_n < \tau\})'$ as before. Let Φ denote the cumulative distribution function of the standard normal.

Lemma 2 (Probability of $\mathbb{A} \cap \mathbb{B}$)**.** *Let $\{U_i : i = 1, \ldots, n\}$ be independent and identically distributed as $\mathbf{N}(0, \sigma^2)$. Then*

$$\mathbb{P}\{\mathbb{A} \cap \mathbb{B}\} \ge 1 - 6M\Phi\left(-\frac{\mu \sqrt{n r_n}}{2\sigma} \lambda \right).$$

Note that r_n depends on the the lower bound t_0 of the parameter space for τ_0. Suppose that t_0 is taken such that $t_0 < \min_{i=1,\ldots,n} Q_i$. Then $\left\| X^{(j)}(t_0) \right\|_n = 0$, and therefore, $r_n = 0$. In this case, Lemma 2 reduces to $\mathbb{P}\{\mathbb{A} \cap \mathbb{B}\} \ge 1 - 3M$ regardless of n and λ, hence resulting in a useless bound. This illustrates a need for restricting the parameter space for τ_0. In practice, researchers tend to choose a sufficiently strict subset of the range of observed values of the threshold variable.

We are ready to establish the prediction consistency of the Lasso estimator. Define $X_{\max} := \max(\mathbf{D})$ and $X_{\min} := \min(\mathbf{D}(t_0))$. Also, let $\alpha_{\max}$ denote the maximum value that all the elements of α can take in absolute value.

Theorem 3 (Consistency of the Lasso)**.** *Let $(\widehat{\alpha}, \widehat{\tau})$ be the Lasso estimator defined by (2.4) with*

$$\lambda = A\sigma\Big(\frac{\log 3M}{nr_n}\Big)^{1/2} \tag{4.4}$$

for a constant $A > 2\sqrt{2}/\mu$ and $r_n > 0$ defined by (4.3). Then, with probability at least $1-(3M)^{1-A^2\mu^2/8}$, we have

$$\left\|\widehat{f}-f_0\right\|_n \leq \Big(6\lambda X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0) + 2\mu\lambda X_{\max}\left|\delta_0\right|_1\Big)^{1/2}.$$

The nonasymptotic upper bound on the prediction loss in Theorem 3 can be easily translated into asymptotic convergence. Specifically, if $X_{\max}$ and $\alpha_{\max}$ are bounded, then Theorem 3 gives

$$\left\|\widehat{f}-f_0\right\|_n \lesssim \sqrt{\lambda\mathcal{M}(\alpha_0)}.$$

Hence, Theorem 3 implies the consistency of the Lasso, provided that $n \to \infty$, $M \to \infty$, and $\lambda\mathcal{M}(\alpha_0) \to 0$. The last condition requires that the sparsity of the model be of smaller order than $\sqrt{(nr_n)/\log 3M}$.

Note that $r_n > 0$ is assumed to be strictly positive in Theorem 3. It requires that $t_0 > \min_{i=1,\dots,n} Q_i$ among other things. We assume that $r_n > 0$ and also $X_{\min} > 0$ throughout the remainder of the paper.

5. Oracle Inequalities

In this section, we establish sparsity oracle inequalities in terms of both the prediction loss and the ℓ_1 estimation loss for unknown parameters. First of all, we make the following assumption.

Assumption 1 (Uniform Restricted Eigenvalue (URE) $(s, c_0, \mathbb{S})$). *For some integer s such that $1 \leq s \leq 2M$ and a positive number c_0, the following condition holds:*

$$\kappa(s, c_0, \mathbb{S}) := \min_{\tau \in \mathbb{S}} \min_{\substack{J_0 \subseteq \{1,\ldots,2M\}, \\ |J_0| \leq s}} \min_{\substack{\gamma \neq 0, \\ \left|\gamma_{J_0^c}\right|_1 \leq c_0 \left|\gamma_{J_0}\right|_1}} \frac{|\mathbf{X}(\tau)\gamma|_2}{\sqrt{n}|\gamma_{J_0}|_2} > 0.$$

If τ_0 were known, then Assumption 1 is just a restatement of the restricted eigenvalue assumption of Bickel et al. (2009) with $\mathbb{S} = \{\tau_0\}$. Bickel et al. (2009) provide sufficient conditions for the restricted eigenvalue condition. In addition, van de Geer and Bühlmann (2009) show the relations between the restricted eigenvalue condition and other conditions on the design matrix.

If τ_0 is unknown as in our setup, it seems necessary to assume that the restricted eigenvalue condition holds uniformly over τ. We consider separately two cases depending on whether $\delta_0 = 0$ or not. On the one hand, if $\delta_0 = 0$ so that τ_0 is not identifiable, then we need to assume that the URE condition holds uniformly on the whole parameter space, $\mathbb{T}$. On the other hand, if $\delta_0 \neq 0$ so that τ_0 is identifiable, then it suffices to impose the URE condition holds uniformly on a neighborhood of τ_0.

We provide a sufficient condition for Assumption 1 below. To that end, we first write $\mathbf{X}(\tau) = (\tilde{\mathbf{X}}, \tilde{\mathbf{X}}(\tau))$ where $\tilde{\mathbf{X}}$ is the $(n \times M)$ matrix whose i-th row is X_i', and $\tilde{\mathbf{X}}(\tau)$ is the $(n \times M)$ matrix whose i-th row is $X_i' 1\{Q_i < \tau\}$, respectively. Define the following Gram matrices:

$$\Psi_n(\tau) := n^{-1}\mathbf{X}(\tau)'\mathbf{X}(\tau),$$
$$\Psi_{n,+}(\tau) := n^{-1}\tilde{\mathbf{X}}(\tau)'\tilde{\mathbf{X}}(\tau),$$
$$\Psi_{n,-}(\tau) := n^{-1}\left[\tilde{\mathbf{X}} - \tilde{\mathbf{X}}(\tau)\right]'\left[\tilde{\mathbf{X}} - \tilde{\mathbf{X}}(\tau)\right],$$

and define the following restricted eigenvalues:

$$\phi_{\min}(u,\tau) := \min_{x\in\mathbb{R}^{2M}:1\le\mathcal{M}(x)\le u}\frac{x'\Psi_n(\tau)x}{x'x},\ \phi_{\max}(u,\tau) := \max_{x\in\mathbb{R}^{2M}:1\le\mathcal{M}(x)\le u}\frac{x'\Psi_n(\tau)x}{x'x},$$
$$\phi_{\min,+}(u,\tau) := \min_{x\in\mathbb{R}^{M}:1\le\mathcal{M}(x)\le u}\frac{x'\Psi_{n,+}(\tau)x}{x'x}, \phi_{\max,+}(u,\tau) := \max_{x\in\mathbb{R}^{M}:1\le\mathcal{M}(x)\le u}\frac{x'\Psi_{n,+}(\tau)x}{x'x}$$

and $\phi_{\min,-}(u,\tau)$ and $\phi_{\max,-}(u,\tau)$ are defined analogously with $\Psi_{n,-}(\tau)$. Let

$$\kappa_2(s,m,c_0,\tau) := \sqrt{\phi_{\min}(s+m,\tau)}\left(1 - c_0\sqrt{\frac{s\phi_{\max}(m,\tau)}{m\phi_{\min}(s+m,\tau)}}\right),$$
$$\psi := \min_{\tau\in\mathbb{S}}\frac{\phi_{\max,-}(2m,\tau)\wedge\phi_{\max,+}(2m,\tau)}{\phi_{\max,-}(2m,\tau)\vee\phi_{\max,+}(2m,\tau)}.$$

Lemma 4. *Assume that the following holds uniformly in* $\tau\in\mathbb{S}$*:*

$$\begin{aligned} m\phi_{\min,+}(2s+2m,\tau) &> c_1^2 s\phi_{\max,+}(2m,\tau), \\ m\phi_{\min,-}(2s+2m,\tau) &> c_1^2 s\phi_{\max,-}(2m,\tau) \end{aligned} \tag{5.1}$$

for some integers s,m*, such that* $1\le s\le M/4$*,* $m\ge s$ *and* $2s+2m\le M$ *and a constant* $c_1>0$*. Also, assume that* $\psi>0$*. Then Assumption 1 is satisfied with* $c_0 = c_1\sqrt{\psi/(1+\psi)}$ *and* $\kappa(s,c_0,\mathbb{S}) = \min_{\tau\in\mathbb{S}}\kappa_2(s,m,c_0,\tau)$*.*

Conditions in (5.1) are modifications of Assumption 2 of Bickel et al. (2009). Note that for each $\tau\in\mathbb{S}$, data are split into two subsamples with corresponding Gram matrices $\Psi_{n,+}(\tau)$ $\Psi_{n,-}(\tau)$, respectively. Hence, conditions in (5.1) are equivalent to stating that Assumption 2 of Bickel et al. (2009) holds with a universal constant c_0 for each subsample of all possible sample splitting induced by different values of $\tau \in \mathbb{S}$. As discussed by Bickel et al. (2009), if we take $s + m = s \log n$ and assume that $\phi_{\max,+}(\cdot,\cdot)$ and $\phi_{\max,-}(\cdot,\cdot)$ are uniformly bounded by a constant, conditions in Lemma 4 are equivalent to

$$\min_{\tau \in \mathbb{S}} \log n \left[\phi_{\min,+}(2s \log n, \tau) \wedge \phi_{\min,-}(2s \log n, \tau)\right] > c_{URE},$$

where $c_{URE} > 0$ is a constant.

The strength of the Lasso method is that it is not necessary to know or pretest whether $\delta_0 = 0$ or not. It is worth noting that we establish oracle inequalities both when $\delta_0 = 0$ and when $\delta_0 \neq 0$. Therefore, the oracle inequalities hold regardless of the existence of threshold effects, implying that we can make prediction without knowing the presence of threshold effect or without pretesting for it.

The following assumption is useful to derive oracle inequalities for each case.

Assumption 2. *Assume that the largest eigenvalue of* $\mathbf{X}(\tau)'\mathbf{X}(\tau)/n$ *is bounded uniformly in* $\tau \in \mathbb{T}$ *by* $\phi_{\max}$.

5.1. **Case I. No Threshold.** We first consider the case that $\delta_0 = 0$. In other words, we estimate a threshold model via the Lasso, but the true model is simply a linear model $Y_i = X_i'\beta_0 + U_i$. This is an important case to consider since in applications, one may not be sure not only about covariates selection but also about the existence of the threshold in the model.

Theorem 5. *Assume that* $\delta_0 = 0$ *and that Assumption 1 holds with* $\kappa = \kappa(s, \frac{1+\mu}{1-\mu}, \mathbb{T})$ *for* $\mu < 1$*, and* $\mathcal{M}(\alpha_0) \leq s \leq M$*. Let* $(\widehat{\alpha}, \widehat{\tau})$ *be the Lasso estimator defined by* (2.4) *with* λ *given by* (4.4)*. Then, with probability at least* $1-(3M)^{1-A^2\mu^2/8}$*, we have*

$$\begin{aligned} \left\|\widehat{f}-f_0\right\|_n &\leq \frac{2A\sigma X_{\max}}{\kappa}\left(\frac{\log 3M}{nr_n}s\right)^{1/2}, \\ \left|\widehat{\alpha}-\alpha_0\right|_1 &\leq \frac{4A\sigma}{(1-\mu)\,\kappa^2}\frac{X_{\max}^2}{X_{\min}}\left(\frac{\log 3M}{nr_n}\right)^{1/2}s. \end{aligned}$$

Furthermore, if Assumption 2 *holds, then*

$$\mathcal{M}(\widehat{\alpha}) \leq \frac{16\phi_{\max}}{(1-\mu)^2\,\kappa^2}\frac{X_{\max}^2}{X_{\min}^2}s.$$

To appreciate the usefulness of the inequalities derived above, it is worth comparing inequalities in Theorem 5 with those in Theorem 7.2 of Bickel et al. (2009). The latter corresponds to the case that $\delta_0 = 0$ is known *a priori*, $\lambda = 2A\sigma(\log M/n)^{1/2}$, $\mu = 1/2$, and $X_{\max} = 1$ using our notation. If we compare Theorem 5 with Theorem 7.2 of Bickel et al. (2009), we can see that the Lasso estimator in (2.4) gives qualitatively the same oracle inequalities as the Lasso estimator in the linear model, even though our model is much more overparametrized in that δ and τ are added to β as parameters to estimate.

5.2. **Case II. Fixed Threshold.** This subsection explores the case where the threshold effect is well-identified and discontinuous. We begin with the following additional assumptions to reflect this.

Assumption 3 (Identifiability under Sparsity and Discontinuity of Regression)**.** *For a given* $s \geq \mathcal{M}(\alpha_0)$*, and for any* η *and* τ *such that* $|\tau-\tau_0| > \eta \geq \min_{i\neq j}|Q_i - Q_j|$ *and* $\alpha \in \{\alpha : \mathcal{M}(\alpha) \leq s\}$*, there exists a* $c > 0$ *such that*

$$\left\|f_{(\alpha,\tau)} - f_0\right\|_n^2 > c\eta.$$

Assumption 3 implies, among other things, that for some $s \geq \mathcal{M}(\alpha_0)$, and for any $\alpha \in \{\alpha : \mathcal{M}(\alpha) \leq s\}$ and τ such that $(\alpha, \tau) \neq (\alpha_0, \tau_0)$,

$$\left\| f_{(\alpha,\tau)} - f_0 \right\|_n \neq 0. \tag{5.2}$$

This condition can be regarded as identifiability of τ_0. If τ_0 were known, then a sufficient condition for the identifiability under the sparsity would be that $URE(s, c_0, \{\tau_0\})$ holds for some $c_0 \geq 1$. Thus, the main point in (5.2) is that there is no sparse representation that is equivalent to f_0 when the sample is split by $\tau \neq \tau_0$. In fact, Assumption 3 is stronger than just the identifiability of τ_0 as it specifies the rate of deviation in f as τ moves away from τ_0, which in turn dictates the convergence rates of $\widehat{\tau}$. We provide further discussions on Assumption 3 in Appendix A.

Remark 1. *The restriction $\eta \geq \min_{i \neq j} |Q_i - Q_j|$ in Assumption 3 is necessary since we consider the fixed design for both X_i and Q_i. Throughout this section, we implicitly assume that the sample size n is large enough such that $\min_{i \neq j} |Q_i - Q_j|$ is very small, implying that the restriction $\eta \geq \min_{i \neq j} |Q_i - Q_j|$ never binds in any of inequalities below. This is typically true for the random design case if Q_i is continuously distributed.*

Assumption 4 (Smoothness of Design)**.** *For any $\eta > 0$, there exists $C < \infty$ such that*

$$\sup_j \sup_{|\tau - \tau_0| < \eta} \frac{1}{n} \sum_{i=1}^{n} \left| X_i^{(j)} \right|^2 \left| 1(Q_i < \tau_0) - 1(Q_i < \tau) \right| \leq C\eta.$$

Assumption 4 has been assumed in the classical setup with a fixed number of stochastic regressors to exclude cases like Q_i has a point mass at τ_0 or $\mathbb{E}(X_i | Q_i = \tau_0)$ is unbounded. In our setup, Assumption 4 amounts to a deterministic version of some smoothness assumption for the distribution of the threshold variable Q_i in the classical setup with stochastic variables. When (X_i, Q_i) is a random vector, it is satisfied under the standard assumption that Q_i is continuously distributed and $\mathbb{E}(|X_i^{(j)}|^2|Q_i = \tau)$ is continuous and bounded in a neighborhood of τ_0 for each j.

To simplify notation, in this section, we assume without loss of generality that $Q_i = i/n$. Then $\mathbb{T} = [t_0, t_1] \subset (0, 1)$. For some constant $\eta > 0$, define an event

$$\mathbb{C}(\eta) := \left\{ \sup_{|\tau-\tau_0|<\eta} \left| \frac{2}{n} \sum_{i=1}^{n} U_i X_i' \delta_0 \left[1(Q_i < \tau_0) - 1(Q_i < \tau) \right] \right| \leq \lambda\sqrt{\eta} \right\},$$

and $h_n(\eta) := \left((2n\eta)^{-1} \sum_{i=[n(\tau_0-\eta)]}^{[n(\tau_0+\eta)]} (X_i'\delta_0)^2 \right)^{1/2}$.

The following lemma gives the lower bound of the probability of the event $\mathbb{A} \cap \mathbb{B} \cap [\cap_{j=1}^{m} \mathbb{C}(\eta_j)]$ for a given m and some positive constants $\eta_1, ..., \eta_m$. To deal with the event $\cap_{j=1}^{m} \mathbb{C}(\eta_j)$, an extra term is added to the lower bound of the probability, in comparison to Lemma 2.

Lemma 6 (Probability of $\mathbb{A} \cap \mathbb{B} \cap \{\cap_{j=1}^{m} \mathbb{C}(\eta_j)\}$)**.** *For a given m and some positive constants $\eta_1, ..., \eta_m$ such that $h_n(\eta_j) > 0$ for each $j = 1, \ldots, m$,*

$$\mathbb{P}\left\{ \mathbb{A} \bigcap \mathbb{B} \bigcap \left[\bigcap_{j=1}^{m} \mathbb{C}(\eta_j) \right] \right\} \geq 1 - 6M\Phi\left(-\frac{\mu\sqrt{nr_n}}{2\sigma}\lambda \right) - 4\sum_{j=1}^{m} \Phi\left(-\frac{\lambda\sqrt{n}}{2\sqrt{2}\sigma h_n(\eta_j)} \right).$$

The following lemma gives an upper bound of $|\widehat{\tau} - \tau_0|$ using only Assumption 3, conditional on the events $\mathbb{A}$ and $\mathbb{B}$.

Lemma 7. *Let $\eta^* = \max\left\{ \min_{i \neq j} |Q_i - Q_j|, c^{-1}\lambda\left(6X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0) + 2\mu X_{\max}|\delta_0|_1\right) \right\}$. Suppose that Assumption 3 holds. Then conditional on the events $\mathbb{A}$ and $\mathbb{B}$,*

$$|\widehat{\tau} - \tau_0| \leq \eta^*.$$

Remark 2. *The nonasymptotic bound in Lemma 7 can be translated into the consistency of $\widehat{\tau}$, as in Theorem 3. That is, if $n \to \infty$, $M \to \infty$, and $\lambda\mathcal{M}(\alpha_0) \to 0$,* *Lemma 7 implies the consistency of* $\widehat{\tau}$*, provided that* $X_{\max}$*,* $\alpha_{\max}$*, and* c^{-1} *are bounded uniformly in* n *and* Q_i *is continuously distributed.*

We now provide a lemma for bounding the prediction risk as well as the ℓ_1 estimation loss for α_0.

Lemma 8. *Suppose that* $|\hat{\tau}-\tau_0| \le c_\tau$ *and* $|\hat{\alpha}-\alpha_0|_1 \le c_\alpha$ *for some* (c_τ, c_α)*. Suppose further that Assumption 4 and Assumption 1 hold with* $\mathbb{S} = \{|\tau-\tau_0| \le c_\tau\}$*,* $\kappa = \kappa(s, \frac{2+\mu}{1-\mu}, \mathbb{S})$ *for* $0<\mu<1$ *and* $\mathcal{M}(\alpha_0) \le s \le M$*. Then, conditional on* $\mathbb{A}$*,* $\mathbb{B}$ *and* $\mathbb{C}(c_\tau)$*, we have*

$$\left\|\widehat{f}-f_0\right\|_n^2 \le 3\lambda\left\{\sqrt{c_\tau}+(2X_{\min})^{-1}c_\tau C\left|\delta_0\right|_1 \vee \frac{6X_{\max}^2}{\kappa^2}\lambda s \vee \frac{2X_{\max}}{\kappa}\left(c_\alpha c_\tau C\left|\delta_0\right|_1 s\right)^{1/2}\right\},$$

$$\left|\widehat{\alpha}-\alpha_0\right|_1 \le \frac{3}{(1-\mu)X_{\min}}\left\{\sqrt{c_\tau}+(2X_{\min})^{-1}c_\tau C\left|\delta_0\right|_1 \vee \frac{6X_{\max}^2}{\kappa^2}\lambda s \vee \frac{2X_{\max}}{\kappa}\left(c_\alpha c_\tau C\left|\delta_0\right|_1 s\right)^{1/2}\right\}.$$

Lemma 8 states the bounds for both $\left\|\widehat{f}-f_0\right\|_n$ and $\left|\widehat{\alpha}-\alpha_0\right|_1$ may become smaller as c_τ gets smaller. This is because decreasing c_τ reduces the first and third terms in the bounds directly, and also because decreasing c_τ reduces the second term in the bound indirectly by allowing for a possibly larger κ since $\mathbb{S}$ gets smaller.

The following lemma shows that the bound for $|\hat{\tau}-\tau_0|$ can be further tightened if we combine results obtained in Lemmas 7 and 8.

Lemma 9. *Suppose that* $|\hat{\tau}-\tau_0| \le c_\tau$ *and* $|\hat{\alpha}-\alpha_0|_1 \le c_\alpha$ *for some* (c_τ, c_α)*. Let* $\tilde{\eta} := c^{-1}\lambda\left((1+\mu)X_{\max}c_\alpha+\sqrt{c_\tau}+(2X_{\min})^{-1}c_\tau C\left|\delta_0\right|_1\right)$*. If Assumption 3 holds, then conditional on the events* $\mathbb{A}$*,* $\mathbb{B}$*, and* $\mathbb{C}(c_\tau)$*,*

$$|\hat{\tau}-\tau_0| \le \tilde{\eta}.$$

Lemma 8 provides us with three different bounds for $|\hat{\alpha}-\alpha_0|_1$ and the two of them are functions of c_τ and c_α. This leads us to apply Lemmas 8 and 9 iteratively to tighten up the bounds. Furthermore, when the sample size is large and thus λ in (4.4) is small enough, we show that the consequence of this chaining argument is that the bound for $|\hat{\alpha}-\alpha_0|$ is dominated by the middle term in Lemma 8. We give exact conditions for this on λ and thus on the sample size n. To do so, we first define some constants:

$$A_{1*} := \frac{3\,(1+\mu)\,X_{\max}}{(1-\mu)\,X_{\min}}+1,\, A_{2*} := \frac{C}{2cX_{\min}},\, A_{3*} := \frac{6cX_{\max}^2}{\kappa^2}, \quad \text{and} \quad A_{4*} := \frac{36\,(1+\mu)\,X_{\max}^3}{(1-\mu)^2X_{\min}}.$$

Assumption 5 (Inequality Conditions)**.** *The following inequalities hold:*

$$A_{1*}A_{2*}\lambda\left|\delta_0\right|_1 < 1, \tag{5.3}$$

$$\frac{A_{1*}}{\left(1-A_{1*}A_{2*}\lambda\left|\delta_0\right|_1\right)^2} < A_{3*}s, \tag{5.4}$$

$$\left(2\kappa^{-2}A_{4*}s+1\right)A_{2*}\lambda\left|\delta_0\right|_1 < 1, \tag{5.5}$$

$$\frac{A_{2*}\lambda\left|\delta_0\right|_1}{\left[1-\left(2\kappa^{-2}A_{4*}s+1\right)A_{2*}\lambda\left|\delta_0\right|_1\right]^2} < \frac{(1-\mu)c}{4}, \tag{5.6}$$

$$\left[1-\left(2\kappa^{-2}A_{4*}s+1\right)A_{2*}\lambda\left|\delta_0\right|_1\right]^{-2} < A_{1*}A_{3*}s. \tag{5.7}$$

Remark 3. *It would be easier to satisfy Assumption 5 when the sample size n is large. To appreciate Assumption 5 in a setup when n is large, suppose that (1) $n\to\infty$, $M\to\infty$, $s\to\infty$, and $\lambda\to 0$; (2) $\left|\delta_0\right|_1$ may or may not diverge to infinity; (3) $X_{\min}$, $X_{\max}$, κ, c, C, and μ are independent of n. Then conditions in Assumption 5 can hold simultaneously for all sufficiently large n, provided that $s\lambda\left|\delta_0\right|_1\to 0$.*

We now give the main result of this section.

Theorem 10. *Suppose that Assumption 1 hold with $\mathbb{S}=\{|\tau-\tau_0|\le\eta^*\}$, $\kappa=\kappa(s,\frac{2+\mu}{1-\mu},\mathbb{S})$ for $0<\mu<1$, and $\mathcal{M}(\alpha_0)\le s\le M$. In addition, Assumptions 4, 3, and 5 hold. Let $(\widehat{\alpha},\widehat{\tau})$ be the Lasso estimator defined by (2.4) with λ given by (4.4). Then, there exist* *a sequence of constants* $\eta_1, ..., \eta_{m^*}$ *for some finite* m^* *such that* $h_n(\eta_j) > 0$ *for each* $j = 1, \ldots, m^*$, *with probability at least* $1-(3M)^{1-A^2\mu^2/8} - 4\sum_{j=1}^{m^*}(3M)^{-A^2/(16r_n h_n(\eta_j))}$, *we have*

$$\left\|\widehat{f} - f_0\right\|_n \leq \frac{3A\sigma X_{\max}}{\kappa}\left(\frac{2\log 3M}{nr_n}s\right)^{1/2},$$

$$|\widehat{\alpha} - \alpha_0|_1 \leq \frac{18A\sigma}{(1-\mu)\kappa^2}\frac{X_{\max}^2}{X_{\min}}\left(\frac{\log 3M}{nr_n}\right)^{1/2}s,$$

and

$$|\hat{\tau} - \tau_0| \leq \left(\frac{3(1+\mu)X_{\max}}{(1-\mu)X_{\min}} + 1\right)\frac{6X_{\max}^2 A^2\sigma^2}{c\kappa^2}\frac{\log 3M}{nr_n}s.$$

Furthermore, if Assumption 2 holds, then

$$\mathcal{M}(\hat{\alpha}) \leq \frac{36\phi_{\max}L_2}{(1-\mu)^2\kappa^2}\frac{X_{\max}^2}{X_{\min}^2}s.$$

Theorem 10 gives the same inequalities (up to constants) as those in Theorem 5 for the prediction risk as well as the ℓ_1 estimation loss for α_0. It is important to note that $|\hat{\tau} - \tau_0|$ is bounded by a constant times $s\log 3M/(nr_n)$, whereas $|\widehat{\alpha} - \alpha_0|_1$ is bounded by a constant times $s[\log 3M/(nr_n)]^{1/2}$. This can be viewed as a nonasymptotic version of the super-consistency of $\widehat{\tau}$ to τ_0. One of main contributions of this paper is that we have extended the well-known super-consistency result of $\widehat{\tau}$ when $M < n$ (see, e.g. Chan, 1993) to the high-dimensional setup ($M \gg n$).

Remark 4. *It is interesting to compare the URE*$(s, c_0, \mathbb{S})$ *condition assumed in Theorem 10 with that in Theorem 5. For the latter, the entire parameter space* $\mathbb{T}$ *is taken to be* $\mathbb{S}$ *but with a smaller constant* $c_0 = (1+\mu)/(1-\mu)$. *Hence, strictly speaking, it is undetermined which URE*$(s, c_0, \mathbb{S})$ *condition is less stringent. It is possible to reduce* c_0 *in Theorem 10 to a smaller constant but larger than* $(1+\mu)/(1-\mu)$ *by considering* *a more general form, e.g.* $c_0 = (1 + \mu + \nu)/(1 - \mu)$ *for a positive constant* ν*, but we have chosen* $\nu = 1$ *here for readability.*

6. Monte Carlo Experiments

In this section we conduct some simulation studies and check the properties of the proposed Lasso estimator. The baseline model is (1.1), where X_i is an M-dimensional vector generated from $N(0, I)$, Q_i is a scalar generated from the uniform distribution in the interval of $(0, 1)$, and the error term U_i is generated from $N(0, 0.5^2)$. The threshold parameter is set to $\tau_0 = 0.3, 0.4$, and 0.5 depending on the simulation design, and the coefficients are set to $\beta_0 = (1, 0, 1, 0, \ldots, 0)$, and $\delta_0 = c \cdot (0, -1, 1, 0, \ldots, 0)$ where $c = 0$ or 1. Note that there is no threshold effect when $c = 0$. The number of observations is set to $n = 200$. Finally, the dimension of X_i in each design is set to $M = 50, 100, 200$ and 400, so that the total number of regressors are 100, 200, 400 and 800, respectively. The range of τ is $\mathbb{T} = [0.15, 0.85]$.

We can estimate the parameters by the standard LASSO/LARS algorithm of Efron et al. (2004) without much modification. Given a regularization parameter value λ, we estimate the model for each grid point of τ that spans over 71 equi-spaced points on $\mathbb{T}$. This procedure can be conducted by using the standard linear Lasso. Next, we plug-in the estimated parameter $\widehat{\alpha}(\tau) := \left(\widehat{\beta}(\tau)', \widehat{\delta}(\tau)'\right)'$ for each τ into the objective function and choose $\widehat{\tau}$ by (2.3). Finally, $\widehat{\alpha}$ is estimated by $\widehat{\alpha}(\widehat{\tau})$. The regularization parameter λ is chosen by (4.4) where $\sigma = 0.5$ is assumed to be known. For the constant A, we use four different values: $A = 2.8, 3.2, 3.6$, and 4.0.

Table 4 and Figures 1–2 summarize these simulation results. To compare the performance of the Lasso estimator, we also report the estimation results of the least squares estimation (Least Squares) available only when $M = 50$ and two oracle models (Oracle 1 and Oracle 2, respectively). Oracle 1 assumes that the regressors with non-zero coefficients are known. In addition to that, Oracle 2 assumes that the true threshold parameter τ_0 is known. Thus, when $c \neq 0$, Oracle 1 estimates $(\beta^{(1)}, \beta^{(3)}, \delta^{(2)}, \delta^{(3)})$ and τ using the least squares estimation while Oracle 2 estimates only $(\beta^{(1)}, \beta^{(3)}, \delta^{(2)}, \delta^{(3)})$. When $c = 0$, both Oracle 1 and Oracle 2 estimate only $(\beta^{(1)}, \beta^{(3)})$. All results are based on 400 replications of each sample.

The reported mean-squared prediction error (PE) for each sample is calculated numerically as follows. For each sample s, we have the estimates $\widehat{\beta}_s$, $\widehat{\delta}_s$, and $\widehat{\tau}_s$. Given these estimates, we generate a new data $\{Y_j, X_j, Q_j\}$ of 400 observations and calculate the prediction error as

$$\widehat{PE}_s = \frac{1}{400} \sum_{j=1}^{400} \Big(f_0(x_j, q_j) - \widehat{f}(x_j, q_j) \Big)^2 . \tag{6.1}$$

The mean, median, and standard deviation of prediction errors are calculated from the 400 replications, $\{\widehat{PE}_s\}_{s=1}^{400}$. In Table 4, we also report mean of $\mathcal{M}(\widehat{\alpha})$ and ℓ_1-errors for α and τ when $M = 50$. For simulation designs with $M > 50$, Least Squares is not available. Figures 1–2 report the same statistics only for the Lasso estimators.

When $M = 50$, across all designs, the proposed Lasso estimator performs better than Least Squares in terms of mean and median prediction errors, mean of $\mathcal{M}(\widehat{\alpha})$, and ℓ_1-error for α. This result becomes more evident when there is no threshold effect, i.e. $c = 0$, which shows the robustness of the Lasso estimator for whether or not there exists a threshold effect. However, the least squares estimator performs better than the Lasso estimator in terms of estimation of τ_0 when $c = 1$, although the difference here is much smaller than the differences in prediction errors and the ℓ_1-error for α.

We can reconfirm the robustness when $M = 100, 200$, and 400 from Figures 1–2. As predicted by the theory developed in previous sections, the prediction errors and ℓ_1 errors for α and τ increase slowly as M increases. The graphs also show that the results are quite uniform across different regularization parameter values except $A = 4.0$.

We next consider different simulation designs. The M-dimensional vector X_i is now generated from a multivariate normal $N(0, \Sigma)$ with $(\Sigma)_{i,j} = \rho^{|i-j|}$, where $(\Sigma)_{i,j}$ denotes the (i,j) element of the $M \times M$ covariance matrix Σ. All other random variables are the same as above. We conducted the simulation studies for both $\rho = 0.1$ and 0.3; however, Tables 5 and Figures 3–4 only report the results of $\rho = 0.3$ to save space (the results with $\rho = 0.1$ are similar). They show very similar results as previous cases: Lasso outperforms Least Squares, and the prediction error, $\mathcal{M}(\widehat{\alpha})$, and ℓ_1-errors increase very slowly as M increases.

Figure 5 shows frequencies of selecting true parameters when both $\rho = 0$ and $\rho = 0.3$. When $\rho = 0$, the probability that the Lasso estimates include the true nonzero parameters is very high. In most cases, the probability is 100%, and even the lowest probability is as high as 98.25%. When $\rho = 0.3$, the corresponding probability is somewhat lower than the no-correlation case, but it is still high and the lowest value is 80.75%.

In sum, the simulation results confirm the theoretical results developed earlier and show that the proposed Lasso estimator will be useful for the threshold model with high-dimensional regressors.

7. Conclusions

We have considered a high-dimensional regression model with a possible change-point due to a covariate threshold and have developed the Lasso method. We have derived nonasymptotic oracle inequalities and have illustrated the usefulness of our proposed estimation method via simulations and a real-data application. It would be an interesting future research topic to extend other penalized estimators (for example, the adaptive Lasso of Zou (2006) and the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001)) to our setup and to see whether we would be able to improve the performance of our estimation method.

Appendix A. Discussions on Assumption 3

We provide further discussions on Assumption 3. Assumption 3 is stronger than just the identifiability of τ_0 as it specifies the rate of deviation in f as τ moves away from τ_0. The linear rate here is sharper than the quadratic one that is usually observed in more regular M-estimation problems, and it reflects the fact that the limit criterion function, in the classical setup with a fixed number of stochastic regressors, has a kink at the true τ_0.

For instance, suppose that $\{(Y_i, X_i, Q_i) : i = 1, \ldots, n\}$ are independent and identically distributed, and consider the case where only the intercept is included in X_i. Assuming that Q_i has a density function that is continuous and positive everywhere (so that $\mathbb{P}(\tau \leq Q_i < \tau_0)$ and $\mathbb{P}(\tau_0 \leq Q_i < \tau)$ can be bounded below by $c_1 |\tau - \tau_0|$ for some $c_1 > 0$), we have that

$$
\begin{aligned}
& \mathbb{E}(Y_i - f_i(\alpha, \tau))^2 - \mathbb{E}(Y_i - f_i(\alpha_0, \tau_0))^2 \\
= \; & \mathbb{E}(f_i(\alpha_0, \tau_0) - f_i(\alpha, \tau))^2 \\
= \; & (\alpha_1 - \alpha_{10})^2 \mathbb{P}(Q_i < \tau \wedge \tau_0) + (\alpha_2 - \alpha_{20})^2 \mathbb{P}(Q_i \geq \tau \vee \tau_0) \\
& + (\alpha_2 - \alpha_{10})^2 \mathbb{P}(\tau \leq Q_i < \tau_0) + (\alpha_1 - \alpha_{20})^2 \mathbb{P}(\tau_0 \leq Q_i < \tau) \\
\geq \; & c |\tau - \tau_0|,
\end{aligned}
$$

for some $c > 0$, where $f_i(\alpha, \tau) = X_i'\beta + X_i'\delta 1\{Q_i < \tau\}$, $\alpha_1 = \beta + \delta$ and $\alpha_2 = \beta$, unless $|\alpha_2 - \alpha_{10}|$ is too small when $\tau < \tau_0$ and $|\alpha_1 - \alpha_{20}|$ is too small when $\tau > \tau_0$. However, when $|\alpha_2 - \alpha_{10}|$ is small, say smaller than ε, $|\alpha_2 - \alpha_{20}|$ is bounded above zero due to the discontinuity that $\alpha_{10} \neq \alpha_{20}$ and $\mathbb{P}(Q_i \geq \tau \vee \tau_0) = \mathbb{P}(Q_i \geq \tau_0)$ is also bounded above zero. This implies the inequality still holds. Since the same reasoning applies for the latter case, we can conclude our discontinuity assumption holds in the standard discontinuous threshold regression setup. In other words, the previous literature has typically imposed conditions sufficient enough to render this condition.

Appendix B. Proofs for Section 4

Proof of Lemma 1. Note that

$$\widehat{S}_n + \lambda \left|\widehat{\mathbf{D}}\widehat{\alpha}\right|_1 \leq S_n(\alpha,\tau) + \lambda \left|\mathbf{D}(\tau)\alpha\right|_1 \tag{B.1}$$

for all $(\alpha,\tau) \in \mathbb{R}^{2M} \times \mathbb{T}$. Now write

$$\begin{aligned}
&\widehat{S}_n - S_n(\alpha,\tau) \\
&= n^{-1}\left|\mathbf{y} - \mathbf{X}(\widehat{\tau})\widehat{\alpha}\right|_2^2 - n^{-1}\left|\mathbf{y} - \mathbf{X}(\tau)\alpha\right|_2^2 \\
&= n^{-1}\sum_{i=1}^n \left[U_i - \{\mathbf{X}_i(\widehat{\tau})'\widehat{\alpha} - \mathbf{X}_i(\tau_0)'\alpha_0\}\right]^2 - n^{-1}\sum_{i=1}^n \left[U_i - \{\mathbf{X}_i(\tau)'\alpha - \mathbf{X}_i(\tau_0)'\alpha_0\}\right]^2 \\
&= n^{-1}\sum_{i=1}^n \{\mathbf{X}_i(\widehat{\tau})'\widehat{\alpha} - \mathbf{X}_i(\tau_0)'\alpha_0\}^2 - n^{-1}\sum_{i=1}^n \{\mathbf{X}_i(\tau)'\alpha - \mathbf{X}_i(\tau_0)'\alpha_0\}^2 \\
&\quad - 2n^{-1}\sum_{i=1}^n U_i\{\mathbf{X}_i(\widehat{\tau})'\widehat{\alpha} - \mathbf{X}_i(\tau)'\alpha\} \\
&= \left\|\widehat{f} - f_0\right\|_n^2 - \left\|f_{(\alpha,\tau)} - f_0\right\|_n^2 \\
&\quad - 2n^{-1}\sum_{i=1}^n U_i X_i'(\widehat{\beta} - \beta) - 2n^{-1}\sum_{i=1}^n U_i\left\{X_i'\widehat{\delta}1(Q_i < \widehat{\tau}) - X_i'\delta 1(Q_i < \tau)\right\}.
\end{aligned}$$

Further, write the last term above as

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n} U_i \left\{ X_i'\widehat{\delta}1(Q_i < \widehat{\tau}) - X_i'\delta 1(Q_i < \tau) \right\} \\
&= n^{-1}\sum_{i=1}^{n} U_i X_i'(\widehat{\delta} - \delta)1(Q_i < \widehat{\tau}) + n^{-1}\sum_{i=1}^{n} U_i X_i'\delta \left\{ 1(Q_i < \widehat{\tau}) - 1(Q_i < \tau) \right\}.
\end{aligned}
$$

Hence, (B.1) can be written as

$$
\begin{aligned}
\left\| \widehat{f} - f_0 \right\|_n^2 \le \left\| f_{(\alpha,\tau)} - f_0 \right\|_n^2 &+ \lambda \left| \mathbf{D}(\tau)\alpha \right|_1 - \lambda \left| \widehat{\mathbf{D}}\widehat{\alpha} \right|_1 \\
&+ 2n^{-1}\sum_{i=1}^{n} U_i X_i'(\widehat{\beta} - \beta) + 2n^{-1}\sum_{i=1}^{n} U_i X_i'(\widehat{\delta} - \delta)1(Q_i < \widehat{\tau}) \\
&+ 2n^{-1}\sum_{i=1}^{n} U_i X_i'\delta \left\{ 1(Q_i < \widehat{\tau}) - 1(Q_i < \tau) \right\}.
\end{aligned}
$$

Then on the events $\mathbb{A}$ and $\mathbb{B}$, we have

$$
\begin{aligned}
\left\| \widehat{f} - f_0 \right\|_n^2 \le \left\| f_{(\alpha,\tau)} - f_0 \right\|_n^2 &+ \mu\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha) \right|_1 \\
&+ \lambda \left| \mathbf{D}(\tau)\alpha \right|_1 - \lambda \left| \widehat{\mathbf{D}}\widehat{\alpha} \right|_1 + R_n(\alpha, \tau)
\end{aligned}
\tag{B.2}
$$

for all $(\alpha, \tau) \in \mathbb{R}^{2M} \times \mathbb{T}$.

Note the the fact that

$$
\left| \widehat{\alpha}^{(j)} - \alpha_0^{(j)} \right| + \left| \alpha_0^{(j)} \right| - \left| \widehat{\alpha}^{(j)} \right| = 0 \text{ for } j \notin J_0.
\tag{B.3}
$$

On the one hand, by (B.2) (evaluating at $(\alpha, \tau) = (\alpha_0, \tau_0)$), on the events $\mathbb{A}$ and $\mathbb{B}$,

$$
\begin{aligned}
&\left\| \widehat{f} - f_0 \right\|_n^2 + (1 - \mu)\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0) \right|_1 \\
&\le \lambda \left(\left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0) \right|_1 + \left| \widehat{\mathbf{D}}\alpha_0 \right|_1 - \left| \widehat{\mathbf{D}}\widehat{\alpha} \right|_1 \right) \\
&+ \lambda \left| \left| \widehat{\mathbf{D}}\alpha_0 \right|_1 - \left| \mathbf{D}\alpha_0 \right|_1 \right| + R_n(\alpha_0, \tau_0) \\
&\le 2\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0} \right|_1 + \lambda \left| \left| \widehat{\mathbf{D}}\alpha_0 \right|_1 - \left| \mathbf{D}\alpha_0 \right|_1 \right| + R_n(\alpha_0, \tau_0),
\end{aligned}
$$

which proves (4.1). On the other hand, again by (B.2) (evaluating at $(\alpha,\tau)=(\alpha_0,\widehat{\tau})$), on the events $\mathbb{A}$ and $\mathbb{B}$,

$$\begin{aligned}
&\left\|\widehat{f}-f_0\right\|_n^2+(1-\mu)\,\lambda\left|\hat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)\right|_1\\
&\le\lambda\left(\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)\right|_1+\left|\widehat{\mathbf{D}}\alpha_0\right|_1-\left|\widehat{\mathbf{D}}\widehat{\alpha}\right|_1\right)+\left\|f_{(\alpha_0,\widehat{\tau})}-f_0\right\|_n^2\\
&\le 2\lambda\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1+\left\|f_{(\alpha_0,\widehat{\tau})}-f_0\right\|_n^2,
\end{aligned}$$

which proves (4.2). □

Proof of Lemma 2. Since $U_i\sim\mathbf{N}(0,\sigma^2)$,

$$\mathbb{P}\{\mathbb{A}^c\}\le\sum_{j=1}^{M}\mathbb{P}\left\{\sqrt{n}|V_{1j}|>\mu\sqrt{n}\lambda/(2\sigma)\right\}=2M\Phi\left(-\frac{\mu\sqrt{n}}{2\sigma}\lambda\right)\le 2M\Phi\left(-\frac{\mu\sqrt{r_n n}}{2\sigma}\lambda\right),$$

where the last inequality follows from $r_n\le 1$.

Now consider the event $\mathbb{B}$. Note that $\left\|X^{(j)}(\tau)\right\|_n$ is monotonically increasing in τ and $\sum_{i=1}^n U_iX_i^{(j)}1\{Q_i<\tau\}$ can be rewritten as a partial sum process by the rearrangement of i according to the magnitude of Q_i. To simplify notation, we assume without loss of generality that $Q_i=i/n$. Then, by Lévy's inequality (see e.g. Proposition A.1.2 of van der Vaart and Wellner, 1996),

$$\begin{aligned}
\mathbb{P}\left\{\sup_{\tau\in\mathbb{T}}\sqrt{n}\,|V_{2j}(\tau)|>\mu\sqrt{n}\lambda/(2\sigma)\right\}&\le\mathbb{P}\left\{\sup_{1\le s\le n}\left|\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{s}U_iX_i^{(j)}\right|>\left\|X^{(j)}\left(t_0\right)\right\|_n\frac{\mu\sqrt{n}}{2\sigma}\lambda\right\}\\
&\le 2\mathbb{P}\left\{\sqrt{n}|V_{1j}|>\frac{\left\|X^{(j)}(t_0)\right\|_n}{\left\|X^{(j)}\right\|_n}\frac{\mu\sqrt{n}}{2\sigma}\lambda\right\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{P}\{\mathbb{B}^c\}&\le\sum_{j=1}^{M}\mathbb{P}\left\{\sup_{\tau\in\mathbb{T}}\sqrt{n}|V_{2j}(\tau)|>\mu\sqrt{n}\lambda/(2\sigma)\right\}\\
&\le 4M\Phi\left(-\frac{\mu\sqrt{r_n n}}{2\sigma}\lambda\right).
\end{aligned}$$

Since $\mathbb{P}\{\mathbb{A} \cap \mathbb{B}\} \geq 1 - \mathbb{P}\{\mathbb{A}^c\} - \mathbb{P}\{\mathbb{B}^c\}$, we have proved the lemma. □

Proof of Theorem 3. Note that

$$R_n = 2n^{-1} \sum_{i=1}^{n} U_i X_i' \delta_0 \left\{ 1(Q_i < \widehat{\tau}) - 1(Q_i < \tau_0) \right\}.$$

Then on the event $\mathbb{B}$,

$$\begin{aligned} |R_n| &\leq 2\mu\lambda \sum_{j=1}^{M} \left\| X^{(j)} \right\|_n |\delta_0^{(j)}| \\ &\leq 2\mu\lambda X_{\max} \left|\delta_0\right|_1 . \end{aligned} \tag{B.4}$$

Then, conditional on $\mathbb{A} \cap \mathbb{B}$, combining (B.4) with (4.1) gives

$$\left\| \widehat{f} - f_0 \right\|_n^2 + (1-\mu)\,\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0) \right|_1 \leq 6\lambda X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0) + 2\mu\lambda X_{\max} \left|\delta_0\right|_1 . \tag{B.5}$$

since

$$\begin{aligned} &|\mathbf{D}(\tau)(\widehat{\alpha} - \alpha_0)_{J_0}|_1 \leq 2X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0), \\ &\left| \left|\widehat{\mathbf{D}}\alpha_0\right|_1 - |\mathbf{D}\alpha_0|_1 \right| \leq 2X_{\max} \left|\alpha_0\right|_1 . \end{aligned}$$

Using the bound that $2\Phi(-x) \leq \exp(-x^2/2)$ for $x > 0$ as in equation (B.4) of Bickel et al. (2009), Lemma 2 with λ given by (4.4) implies that the event $\mathbb{A} \cap \mathbb{B}$ occurs with probability at least $1 - (3M)^{1-A^2\mu^2/8}$. Then the theorem follows from (B.5). □

Appendix C. Proofs for Section 5

We first provide a lemma to derive an oracle inequality regarding the sparsity of the Lasso estimator $\widehat{\alpha}$.

Lemma 11 (Sparsity of the Lasso)**.** *Let Assumption 2 hold. Then conditional on the event $\mathbb{A} \bigcap \mathbb{B}$, we have*

$$\mathcal{M}(\widehat{\alpha}) \leq \frac{4\phi_{\max}}{(1-\mu)^2 \lambda^2 X_{\min}^2} \left\| \widehat{f} - f_0 \right\|_n^2 . \tag{C.1}$$

Proof of Lemma 11*.* As in (B.6) of Bickel et al. (2009), for each τ, the necessary and sufficient condition for $\widehat{\alpha}(\tau)$ to be the Lasso solution can be written in the form

$$\begin{aligned}
&\frac{2}{n}[X^{(j)}]'(\mathbf{y} - \mathbf{X}(\tau)\widehat{\alpha}(\tau)) = \lambda \left\| X^{(j)} \right\|_n \operatorname{sign}(\widehat{\beta}^{(j)}(\tau)) && \text{if } \widehat{\beta}^{(j)}(\tau) \neq 0 \\
&\left| \frac{2}{n}[X^{(j)}]'(\mathbf{y} - \mathbf{X}(\tau)\widehat{\alpha}(\tau)) \right| \leq \lambda \left\| X^{(j)} \right\|_n && \text{if } \widehat{\beta}^{(j)}(\tau) = 0 \\
&\frac{2}{n}[X^{(j)}(\tau)]'(\mathbf{y} - \mathbf{X}(\tau)\widehat{\alpha}(\tau)) = \lambda \left\| X^{(j)}(\tau) \right\|_n \operatorname{sign}(\widehat{\delta}^{(j)}(\tau)) && \text{if } \widehat{\delta}^{(j)}(\tau) \neq 0 \\
&\left| \frac{2}{n}[X^{(j)}(\tau)]'(\mathbf{y} - \mathbf{X}(\tau)\widehat{\alpha}(\tau)) \right| \leq \lambda \left\| X^{(j)}(\tau) \right\|_n && \text{if } \widehat{\delta}^{(j)}(\tau) = 0,
\end{aligned}$$

where $j = 1, \ldots, M$.

Note that conditional on events $\mathbb{A}$ and $\mathbb{B}$,

$$\begin{aligned}
\left| \frac{2}{n} \sum_{i=1}^{n} U_i X_i^{(j)} \right| &\leq \mu\lambda \left\| X^{(j)} \right\|_n \\
\left| \frac{2}{n} \sum_{i=1}^{n} U_i X_i^{(j)} 1\{Q_i < \tau\} \right| &\leq \mu\lambda \left\| X^{(j)}(\tau) \right\|_n
\end{aligned}$$

for any τ, where $j = 1, \ldots, M$. Therefore,

$$\begin{aligned}
&\left| \frac{2}{n}[X^{(j)}]'(\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\tau)\widehat{\alpha}(\tau)) \right| \geq (1-\mu)\lambda \left\| X^{(j)} \right\|_n && \text{if } \widehat{\beta}^{(j)}(\tau) \neq 0 \\
&\left| \frac{2}{n}[X^{(j)}(\tau)]'(\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\tau)\widehat{\alpha}(\tau)) \right| \geq (1-\mu)\lambda \left\| X^{(j)}(\tau) \right\|_n && \text{if } \widehat{\delta}^{(j)}(\tau) \neq 0.
\end{aligned}$$

Using inequalities above, write

$$\begin{aligned}
&\frac{1}{n^2}\left[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}\right]' \mathbf{X}(\widehat{\tau})\mathbf{X}(\widehat{\tau})' \left[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}\right] \\
&= \frac{1}{n^2}\sum_{j=1}^{M}\left\{[X^{(j)}]'[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}]\right\}^2 + \frac{1}{n^2}\sum_{j=1}^{M}\left\{[X^{(j)}(\widehat{\tau})]'[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}]\right\}^2 \\
&\geq \frac{1}{n^2}\sum_{j:\widehat{\beta}^{(j)}\neq 0}\left\{[X^{(j)}]'[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}]\right\}^2 + \frac{1}{n^2}\sum_{j:\widehat{\delta}^{(j)}\neq 0}\left\{[X^{(j)}(\widehat{\tau})]'[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}]\right\}^2 \\
&\geq \frac{(1-\mu)^2\lambda^2}{4}\left(\sum_{j:\widehat{\beta}^{(j)}\neq 0}\left\|X^{(j)}\right\|_n^2 + \sum_{j:\widehat{\delta}^{(j)}\neq 0}\left\|X^{(j)}(\widehat{\tau})\right\|_n^2\right) \\
&\geq \frac{(1-\mu)^2\lambda^2}{4} X_{\min}^2 \mathcal{M}\left(\hat{\alpha}\right).
\end{aligned}$$

To complete the proof, note that

$$\begin{aligned}
&\frac{1}{n^2}\left[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}\right]' \mathbf{X}(\widehat{\tau})\mathbf{X}(\widehat{\tau})' \left[\mathbf{X}(\tau_0)\alpha_0 - \mathbf{X}(\widehat{\tau})\widehat{\alpha}\right] \\
&\qquad \leq \operatorname{maxeig}(\mathbf{X}(\widehat{\tau})\mathbf{X}(\widehat{\tau})'/n)\left\|\widehat{f} - f_0\right\|_n^2 \\
&\qquad \leq \phi_{\max}\left\|\widehat{f} - f_0\right\|_n^2,
\end{aligned}$$

where $\operatorname{maxeig}(\mathbf{X}(\widehat{\tau})\mathbf{X}(\widehat{\tau})'/n)$ denotes the largest eigenvalue of $\mathbf{X}(\widehat{\tau})\mathbf{X}(\widehat{\tau})'/n$. □

Proof of Lemma 4. Define $\bar{\mathbf{X}}(\tau) := (\tilde{\mathbf{X}} - \tilde{\mathbf{X}}(\tau), \tilde{\mathbf{X}}(\tau))$. For any $y = (y_1', y_2')'$ such that $y_1, y_2 \in \mathbb{R}^M \setminus \{0\}$, let $x_1 = y_1/\sqrt{y'y}, x_2 = y_2/\sqrt{y'y}$. Then $x_1'x_1 + x_2'x_2 = 1$. Furthermore, since $[\tilde{\mathbf{X}} - \tilde{\mathbf{X}}(\tau)]'\tilde{\mathbf{X}}(\tau) = 0$, we have

$$\frac{y'n^{-1}\bar{\mathbf{X}}(\tau)'\bar{\mathbf{X}}(\tau)y}{y'y} = \frac{x_1'\Psi_{n,-}(\tau)x_1}{x_1'x_1}x_1'x_1 + \frac{x_2'\Psi_{n,+}(\tau)x_2}{x_2'x_2}x_2'x_2.$$

Also, note that $\mathcal{M}(x_1)$ and $\mathcal{M}(x_2)$ are smaller than or equal to $\mathcal{M}(y)$.

Since any selection of s column vectors in $\mathbf{X}(\tau)$ can be represented by a linear transformation of a selection of $2s$ column vectors of $\bar{\mathbf{X}}(\tau)$, the minimum restricted eigenvalue of dimension $2s$ for $\bar{\mathbf{X}}(\tau)$ can be smaller than that of dimension s for $\mathbf{X}(\tau)$. Likewise, the maximum restricted eigenvalue of dimension $2s$ for $\bar{\mathbf{X}}(\tau)$ can be larger than that of dimension s for $\mathbf{X}(\tau)$. Thus, with $u = 2s + 2m$,

$$\begin{aligned} m \min_{y \in \mathbb{R}^{2M}: 1 \leq \mathcal{M}(y) \leq s+m} \frac{y' n^{-1} \mathbf{X}(\tau)' \mathbf{X}(\tau) y}{y'y} &\geq m \min_{y \in \mathbb{R}^{2M}: 1 \leq \mathcal{M}(y) \leq u} \frac{y' n^{-1} \bar{\mathbf{X}}(\tau)' \bar{\mathbf{X}}(\tau) y}{y'y} \\ &\geq m\left(\phi_{\min,-}(u,\tau) \wedge \phi_{\min,+}(u,\tau)\right) \\ &> c_1^2 s\left(\phi_{\max,-}(2m,\tau) \wedge \phi_{\max,+}(2m,\tau)\right) \\ &> c_1^2 s \frac{\psi}{1+\psi} \max_{y \in \mathbb{R}^{2M}: 1 \leq \mathcal{M}(y) \leq 2m} \frac{y' n^{-1} \bar{\mathbf{X}}(\tau)' \bar{\mathbf{X}}(\tau) y}{y'y} \\ &\geq c_1^2 s \frac{\psi}{1+\psi} \max_{y \in \mathbb{R}^{2M}: 1 \leq \mathcal{M}(y) \leq m} \frac{y' n^{-1} \mathbf{X}(\tau)' \mathbf{X}(\tau) y}{y'y}. \end{aligned}$$

This implies that Bickel et al. (2009)'s Assumption 2 hold for $\mathbf{X}(\tau)$ with $c_0 = c_1\sqrt{\psi/(1+\psi)}$. Then, it follows from their Lemma 4.1 that Assumption 1 is satisfied with $\kappa(s, c_0) = \min_{\tau \in \mathbb{S}} \kappa_2(s, m, c_0, \tau)$. □

Proof of Theorem 5. Note that $\delta_0 = 0$ implies $\left\| f_{(\alpha_0, \widehat{\tau})} - f_0 \right\|^2 = 0$. Combining this with (4.2), we have

$$\left\| \widehat{f} - f_0 \right\|_n^2 + (1-\mu)\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0) \right|_1 \leq 2\lambda \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0} \right|_1, \tag{C.2}$$

which implies that

$$\left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0^c} \right|_1 \leq \frac{1+\mu}{1-\mu} \left| \widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0} \right|_1.$$

This in turn allows us to apply Assumption 1, specifically URE$(s, \frac{1+\mu}{1-\mu}, \mathbb{T})$, to yield

$$
\begin{aligned}
\kappa^2 \left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_2^2 &\leq \frac{1}{n}|\mathbf{X}(\widehat{\tau})\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)|_2^2 \\
&= \frac{1}{n}(\widehat{\alpha}-\alpha_0)'\widehat{\mathbf{D}}\mathbf{X}(\widehat{\tau})'\mathbf{X}(\widehat{\tau})\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0) \\
&\leq \frac{\max(\widehat{\mathbf{D}})^2}{n}(\widehat{\alpha}-\alpha_0)'\mathbf{X}(\widehat{\tau})'\mathbf{X}(\widehat{\tau})(\widehat{\alpha}-\alpha_0) \\
&= \max(\widehat{\mathbf{D}})^2 \left\|\widehat{f}-f_0\right\|_n^2,
\end{aligned} \tag{C.3}
$$

where $\kappa = \kappa(s, \frac{1+\mu}{1-\mu}, \mathbb{T})$ and the last equality is due to the assumption that $\delta_0 = 0$.

Combining (C.2) with (C.3) yields

$$
\begin{aligned}
\left\|\widehat{f}-f_0\right\|_n^2 &\leq 2\lambda \left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1 \\
&\leq 2\lambda\sqrt{s}\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_2 \\
&\leq \frac{2\lambda}{\kappa}\sqrt{s}\max(\widehat{\mathbf{D}})\left\|\widehat{f}-f_0\right\|_n .
\end{aligned}
$$

Then the first conclusion of the theorem follows immediately.

In addition, combining the arguments above with the first conclusion of the theorem yields

$$
\begin{aligned}
\left|\widehat{\mathbf{D}}\left(\widehat{\alpha}-\alpha_0\right)\right|_1 &= \left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1 + \left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0^c}\right|_1 \\
&\leq 2\left(1-\mu\right)^{-1}\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1 \\
&\leq 2\left(1-\mu\right)^{-1}\sqrt{s}\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_2 \\
&\leq \frac{2}{\kappa\left(1-\mu\right)}\sqrt{s}\max(\widehat{\mathbf{D}})\left\|\widehat{f}-f_0\right\|_n \\
&\leq \frac{4\lambda}{\left(1-\mu\right)\kappa^2} s X_{\max}^2,
\end{aligned} \tag{C.4}
$$

which proves the second conclusion of the theorem since

$$\left|\widehat{\mathbf{D}}\left(\widehat{\alpha}-\alpha_0\right)\right|_1 \geq \min(\widehat{\mathbf{D}})\left|\widehat{\alpha}-\alpha_0\right|_1. \tag{C.5}$$

Finally, theorem follows by Lemma 11 with the bound on $\mathbb{P}(\mathbb{A}\cap\mathbb{B})$ as in the proof of Theorem 3. □

Proof of Lemma 6. Given Lemma 2, it remains to examine the probability of $\mathbb{C}\left(\eta_j\right)$. As in the proof of Lemma 2, Lévy's inequality yields that

$$\begin{aligned}
\mathbb{P}\left\{\mathbb{C}\left(\eta_j\right)^c\right\} &\leq \mathbb{P}\left\{\sup_{|\tau-\tau_0|\leq\eta_j}\left|\frac{2}{n}\sum_{i=1}^{n}U_iX_i'\delta_0\left[1\left(Q_i<\tau_0\right)-1\left(Q_i<\tau\right)\right]\right|>\lambda\sqrt{\eta_j}\right\}. \\
&\leq 2\mathbb{P}\left\{\left|\frac{2}{n}\sum_{i=[n(\tau_0-\eta_j)]}^{[n(\tau_0+\eta_j)]}U_iX_i'\delta_0\right|>\lambda\sqrt{\eta_j}\right\} \\
&\leq 4\Phi\left(-\frac{\lambda\sqrt{n}}{2\sqrt{2}\sigma h_n\left(\eta_j\right)}\right).
\end{aligned}$$

Hence, we have proved the lemma since $\mathbb{P}\left\{\mathbb{A}\bigcap\mathbb{B}\bigcap\left[\bigcap_{j=1}^{m}\mathbb{C}\left(\eta_j\right)\right]\right\}\geq 1-\mathbb{P}\{\mathbb{A}^c\}-\mathbb{P}\{\mathbb{B}^c\}-\sum_{j=1}^{m}\mathbb{P}\{\mathbb{C}\left(\eta_j\right)^c\}$. □

Proof of Lemma 7. As in the proof of Lemma 1, we have, on the events $\mathbb{A}$ and $\mathbb{B}$,

(C.6)

$$\begin{aligned}
&\widehat{S}_n - S_n(\alpha_0,\tau_0) \\
&= \left\|\widehat{f}-f_0\right\|_n^2 - 2n^{-1}\sum_{i=1}^{n}U_iX_i'(\widehat{\beta}-\beta_0) - 2n^{-1}\sum_{i=1}^{n}U_iX_i'(\widehat{\delta}-\delta_0)1(Q_i<\widehat{\tau}) - R_n \\
&\geq \left\|\widehat{f}-f_0\right\|_n^2 - \mu\lambda\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)\right|_1 - R_n.
\end{aligned}$$

Then using (B.3), on the events $\mathbb{A}$ and $\mathbb{B}$,

$$
\begin{aligned}
&\left[\widehat{S}_n + \lambda \left|\widehat{\mathbf{D}}\widehat{\alpha}\right|_1\right] - \left[S_n(\alpha_0, \tau_0) + \lambda \left|\mathbf{D}\alpha_0\right|_1\right] \\
&\geq \left\|\widehat{f} - f_0\right\|_n^2 - \mu\lambda \left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)\right|_1 - \lambda \left[\left|\mathbf{D}\alpha_0\right|_1 - \left|\widehat{\mathbf{D}}\widehat{\alpha}\right|_1\right] - R_n \\
&\geq \left\|\widehat{f} - f_0\right\|_n^2 - 2\lambda \left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0}\right|_1 - \lambda \left[\left|\mathbf{D}\alpha_0\right|_1 - \left|\widehat{\mathbf{D}}\alpha_0\right|_1\right] - R_n \\
&\geq \left\|\widehat{f} - f_0\right\|_n^2 - \left[6\lambda X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0) + 2\mu\lambda X_{\max}\left|\delta_0\right|_1\right],
\end{aligned}
\tag{C.7}
$$

where the second inequality follows from (B.4) and the last inequality comes from following bounds:

$$
\begin{aligned}
2\lambda \left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0}\right|_1 &\leq 4\lambda X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0), \\
\lambda \left|\left|\mathbf{D}\alpha_0\right|_1 - \left|\widehat{\mathbf{D}}\alpha_0\right|_1\right| &\leq 2\lambda X_{\max}\alpha_{\max}\mathcal{M}(\alpha_0).
\end{aligned}
$$

Suppose now that $|\hat{\tau} - \tau_0| > \eta^*$. Then Assumption 3 and (C.7) together imply that

$$
\left[\widehat{S}_n + \lambda \left|\widehat{\mathbf{D}}\widehat{\alpha}\right|_1\right] - \left[S_n(\alpha_0, \tau_0) + \lambda \left|\mathbf{D}\alpha_0\right|_1\right] > 0,
$$

which leads to contradiction as $\widehat{\tau}$ is the minimizer of the criterion function as in (2.4). Therefore, we have proved the lemma. □

Proof of Lemma 8. Note that on $\mathbb{C}$,

$$
\begin{aligned}
|R_n| &= \left|2n^{-1}\sum_{i=1}^{n} U_i X_i' \delta_0 \left\{1(Q_i < \widehat{\tau}) - 1(Q_i < \tau_0)\right\}\right| \\
&\leq \lambda\sqrt{c_\tau}.
\end{aligned}
$$

The triangular inequality, the mean value theorem (applied to $f(x) = \sqrt{x}$), and Assumption 4 imply that

(C.8)

$$\begin{aligned}\left|\left|\widehat{\mathbf{D}}\alpha_0\right|_1 - |\mathbf{D}\alpha_0|_1\right| &= \left|\sum_{j=1}^{M}\left(\left\|X^{(j)}(\hat{\tau})\right\|_n - \left\|X^{(j)(\tau_0)}\right\|_n\right)\left|\delta_0^{(j)}\right|\right| \\ &\leq \sum_{j=1}^{M}\left(2\left\|X^{(j)}(t_0)\right\|_n\right)^{-1}\left|\delta_0^{(j)}\right|\frac{1}{n}\sum_{i=1}^{n}\left|X_i^{(j)}\right|^2\left|1\{Q_i<\hat{\tau}\}-1\{Q_i<\tau_0\}\right| \\ &\leq (2X_{\min})^{-1}c_\tau C|\delta_0|_1.\end{aligned}$$

We now consider two cases: (i) $\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1 > \sqrt{c_\tau} + (2X_{\min})^{-1}c_\tau C|\delta_0|_1$ and (ii) $\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1 \leq \sqrt{c_\tau} + (2X_{\min})^{-1}c_\tau C|\delta_0|_1$.

Case (i): In this case, note that

$$\begin{aligned}\lambda\left|\left|\widehat{\mathbf{D}}\alpha_0\right|_1 - |\mathbf{D}\alpha_0|_1\right| + R_n &< \lambda(2X_{\min})^{-1}c_\tau C|\delta_0|_1 + \lambda\sqrt{c_\tau} \\ &= \lambda\left(\sqrt{c_\tau} + (2X_{\min})^{-1}c_\tau C|\delta_0|_1\right) \\ &< \lambda\left|\widehat{D}(\widehat{\alpha}-\alpha_0)_{J_0}\right|.\end{aligned}$$

Combining this result with (4.1), we have

$$\left\|\widehat{f}-f_0\right\|_n^2 + (1-\mu)\lambda\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)\right|_1 \leq 3\lambda\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1, \tag{C.9}$$

which implies

$$\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0^c}\right|_1 \leq \frac{2+\mu}{1-\mu}\left|\widehat{\mathbf{D}}(\widehat{\alpha}-\alpha_0)_{J_0}\right|_1.$$

Then, we apply Assumption 1 with URE$(s, \frac{2+\mu}{1-\mu}, \mathbb{S})$. Note that since it is assumed that $|\widehat{\tau}-\tau_0| \leq c_\tau$, Assumption 1 only needs to hold with $\mathbb{S}$ in the c_τ neighborhood of

τ_0. Since $\delta_0 \neq 0$, (C.3) now has an extra term

$$
\begin{aligned}
&\kappa^2 \left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)_{J_0}\right|_2^2 \\
&\leq \max(\widehat{\mathbf{D}})^2 \left\|\hat{f} - f_0\right\|_n^2 \\
&\quad + \max(\widehat{\mathbf{D}})^2 \frac{1}{n} \sum_{i=1}^{n} \left\{ 2\Big(\mathbf{X}_i(\widehat{\tau})'\widehat{\alpha} - \mathbf{X}_i(\widehat{\tau})'\alpha_0\Big)\Big(X_i'\delta_0\big[1(Q_i < \tau_0) - 1(Q_i < \widehat{\tau})\big]\Big) \right\} \\
&\leq \max(\widehat{\mathbf{D}})^2 \left(\left\|\hat{f} - f_0\right\|_n^2 + 2c_\alpha \left|\delta_0\right|_1 \sup_j \frac{1}{n} \sum_{i=1}^{n} \left|X_i^{(j)}\right|^2 \left|1(Q_i < \tau_0) - 1(Q_i < \widehat{\tau})\right| \right) \\
&\leq X_{\max}^2 \bigg(\left\|\widehat{f} - f_0\right\|_n^2 + 2c_\alpha c_\tau C \left|\delta_0\right|_1 \bigg),
\end{aligned}
$$

where the last inequality is due to Assumption 4. Combining this result with (C.9), we have

$$
\begin{aligned}
\left\|\widehat{f} - f_0\right\|_n^2 &\leq 3\lambda \left|\widehat{\mathbf{D}}\left(\widehat{\alpha} - \alpha_0\right)_{J_0}\right|_1 \\
&\leq 3\lambda\sqrt{s} \left|\widehat{\mathbf{D}}\left(\widehat{\alpha} - \alpha_0\right)_{J_0}\right|_2 \\
&\leq 3\lambda\sqrt{s} \left(\kappa^{-2} X_{\max}^2 \left(\left\|\widehat{f} - f_0\right\|_n^2 + 2c_\alpha c_\tau C \left|\delta_0\right|_1 \right) \right)^{1/2}.
\end{aligned}
$$

Applying $a + b \leq 2a \vee 2b$, we get the upper bound of $\left\|\widehat{f} - f_0\right\|_n$ on $\mathbb{A}$ and $\mathbb{B}$, as

$$
\left\|\widehat{f} - f_0\right\|_n^2 \leq \frac{18X_{\max}^2}{\kappa^2}\lambda^2 s \vee \frac{6X_{\max}}{\kappa}\lambda\left(c_\alpha c_\tau C \left|\delta_0\right|_1 s\right)^{1/2}. \tag{C.10}
$$

To derive the upper bound for $|\widehat{\alpha}-\alpha_0|_1$, note that using the the same arguments as in (C.4),

$$\begin{aligned}\left|\widehat{D}(\widehat{\alpha}-\alpha_0)\right|_1 &\leq \frac{3}{1-\mu}\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_{J_0}\right|_1 \\ &\leq \frac{3}{1-\mu}\sqrt{s}\left|\widehat{D}\left(\widehat{\alpha}-\alpha_0\right)_{J_0}\right|_2 \\ &\leq \frac{3}{1-\mu}\sqrt{s}\left(\kappa^{-2}X_{\max}^2\left(\left\|\widehat{f}-f_0\right\|_n^2+2c_\alpha c_\tau C\left|\delta_0\right|_1\right)\right)^{1/2} \\ &\leq \frac{3\sqrt{s}}{(1-\mu)\kappa}X_{\max}\left(\left\|\widehat{f}-f_0\right\|_n^2+2c_\alpha c_\tau C\left|\delta_0\right|_1\right)^{1/2}.\end{aligned}$$

Then combining the fact that $a+b\leq 2a\vee 2b$ with (C.5) and (C.10) yields

$$|\widehat{\alpha}-\alpha_0|_1 \leq \frac{18}{(1-\mu)\kappa^2}\frac{X_{\max}^2}{X_{\min}}\lambda s \vee \frac{6}{(1-\mu)\kappa}\frac{X_{\max}}{X_{\min}}\left(c_\alpha c_\tau C\left|\delta_0\right|_1 s\right)^{1/2}.$$

Case (ii): In this case, it follows directly from (4.1) that

$$\begin{aligned}\left\|\widehat{f}-f_0\right\|_n^2 &\leq 3\lambda\left(\sqrt{c_\tau}+(2X_{\min})^{-1}c_\tau C\left|\delta_0\right|_1\right), \\ |\widehat{\alpha}-\alpha_0|_1 &\leq \frac{3}{(1-\mu)X_{\min}}\left(\sqrt{c_\tau}+(2X_{\min})^{-1}c_\tau C\left|\delta_0\right|_1\right),\end{aligned}$$

which establishes the desired result. □

Proof of Lemma 9. Note that on $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$,

$$\begin{aligned}&\left|\frac{2}{n}\sum_{i=1}^n\left[U_iX_i'\left(\hat{\beta}-\beta_0\right)+U_iX_i'1\left(Q_i<\hat{\tau}\right)\left(\hat{\delta}-\delta_0\right)\right]\right| \\ \leq\ & \mu\lambda X_{\max}\left|\hat{\alpha}-\alpha_0\right|_1 \leq \mu\lambda X_{\max}c_\alpha\end{aligned}$$

and

$$\left|\frac{2}{n}\sum_{i=1}^n U_iX_i'\delta_0\left[1\left(Q_i<\hat{\tau}\right)-1\left(Q_i<\tau_0\right)\right]\right| \leq \lambda\sqrt{c_\tau}.$$

Suppose $\tilde{\eta} < |\hat{\tau} - \tau_0| < c_\tau$. Then, as in (C.6),

$$\widehat{S}_n - S_n(\alpha_0, \tau_0) \geq \left\| \widehat{f} - f_0 \right\|_n^2 - \mu\lambda X_{\max} c_\alpha - \lambda\sqrt{c_\tau}.$$

Furthermore, we obtain

$$\begin{aligned}
&\left[\widehat{S}_n + \lambda \left|\widehat{\mathbf{D}}\widehat{\alpha}\right|_1\right] - \left[S_n(\alpha_0, \tau_0) + \lambda \left|\mathbf{D}\alpha_0\right|_1\right] \\
&\geq \left\| \widehat{f} - f_0 \right\|_n^2 - \mu\lambda X_{\max} c_\alpha - \lambda\sqrt{c_\tau} \\
&- \lambda \left(\left|\widehat{\mathbf{D}}(\widehat{\alpha} - \alpha_0)\right|_1 + \left|(\widehat{\mathbf{D}} - \mathbf{D})\alpha_0\right|_1 \right) \\
&> c\tilde{\eta} - \left((1+\mu) X_{\max} c_\alpha + \sqrt{c_\tau} + (2X_{\min})^{-1} c_\tau C \left|\delta_0\right|_1 \right) \lambda,
\end{aligned}$$

where the last inequality is due to Assumption 3 and (C.8).

Since $c\tilde{\eta} = \left((1+\mu) X_{\max} c_\alpha + \sqrt{c_\tau} + (2X_{\min})^{-1} c_\tau C \left|\delta_0\right|_1 \right) \lambda$, we again use the contradiction argument as in the proof of Lemma 7 to establish the result. □

Proof of of Theorem 10. Here we use the chaining argument by iteratively applying Lemmas 8 and 9 to tighten the bounds for the prediction risk and the estimation errors in $\hat{\alpha}$ and $\hat{\tau}$.

Let c_α^* and c_τ^* denote the bounds given in the main theorem for $|\hat{\alpha} - \alpha_0|_1$ and $|\hat{\tau} - \tau_0|$, respectively. Suppose that

$$\sqrt{c_\tau} + (2X_{\min})^{-1} c_\tau C \left|\delta_0\right|_1 \vee \frac{6X_{\max}^2}{\kappa^2}\lambda s \vee \frac{2X_{\max}}{\kappa} \left(c_\alpha c_\tau C \left|\delta_0\right|_1 s\right)^{1/2} = \frac{6X_{\max}^2}{\kappa^2}\lambda s. \tag{C.11}$$

This implies due to Lemma 8 that $|\hat{\alpha} - \alpha_0|_1$ is bounded by c_α^* and thus achieves the bounds in the theorem given the choice of λ. The same argument applies for $\left\| \hat{f} - f_0 \right\|_n^2$. The equation (C.11) also implies in conjunction with Lemma 9 with $c_\alpha =$

c_α^* that

$$\begin{aligned} |\hat{\tau} - \tau_0| &\leq c^{-1}\lambda\left(\left(1+\mu\right)X_{\max}c_\alpha^* + \sqrt{c_\tau} + \left(2X_{\min}\right)^{-1}c_\tau C\left|\delta_0\right|_1\right) \\ &\leq \left(\frac{3\left(1+\mu\right)X_{\max}}{(1-\mu)X_{\min}} + 1\right)\frac{6X_{\max}^2}{c\kappa^2}\lambda^2 s, \end{aligned} \tag{C.12}$$

which is c_τ^*. Thus, it remains to show that there is convergence in the iterated applications of Lemmas 8 and 9 toward the desired bounds when (C.11) does not hold and the number of iteration is finite.

Let $c_\tau^{(m)}$ and $c_\alpha^{(m)}$, respectively, denote the bounds for $|\hat{\alpha} - \alpha_0|_1$ and $|\hat{\tau} - \tau_0|$ in the m-th iteration. In view of (B.5) and Lemma 7, we start the iteration with

$$c_\alpha^{(1)} := \frac{8X_{\max}\alpha_{\max}}{\left(1-\mu\right)X_{\min}}s,$$
$$c_\tau^{(1)} := c^{-1}8X_{\max}\alpha_{\max}\lambda s.$$

If the starting values $c_\alpha^{(1)}$ and $c_\tau^{(1)}$ are smaller than the desired bounds, we do not start the iteration. Otherwise, we stop the iteration as soon as updated bounds are smaller than the desired bounds.

Since Lemma 8 provides us with two types of bounds for c_α when (C.11) is not met, we evaluate each case below.

Case (i):

$$c_\alpha^{(m)} = \frac{3}{(1-\mu)X_{\min}}\left(\sqrt{c_\tau^{(m-1)}} + \left(2X_{\min}\right)^{-1}c_\tau^{(m-1)}C\left|\delta_0\right|_1\right).$$

This implies by Lemma 9 that

$$
\begin{aligned}
c_\tau^{(m)} &= c^{-1}\lambda\left((1+\mu)\,X_{\max}c_\alpha^{(m)} + \sqrt{c_\tau^{(m-1)}} + (2X_{\min})^{-1}c_\tau^{(m-1)}C\,|\delta_0|_1\right)\\
&= c^{-1}\lambda\left(\frac{3\,(1+\mu)\,X_{\max}}{(1-\mu)\,X_{\min}}+1\right)\left(\sqrt{c_\tau^{(m-1)}} + (2X_{\min})^{-1}\,C\,|\delta_0|_1\,c_\tau^{(m-1)}\right)\\
&=: A_1\sqrt{c_\tau^{(m-1)}} + A_2 c_\tau^{(m-1)},
\end{aligned}
$$

where A_1 and A_2 are defined accordingly. This system has one converging fixed point other than zero if $A_2 < 1$, which is the case under (5.3) . Note also that all the terms here are positive. After some algebra, we get the fixed point

$$
\begin{aligned}
c_\tau^{\infty} &= \left(\frac{A_1}{1-A_2}\right)^2\\
&= \left(\frac{c^{-1}\lambda\left(\frac{3(1+\mu)X_{\max}}{(1-\mu)X_{\min}}+1\right)}{1-c^{-1}\lambda\left(\frac{3(1+\mu)X_{\max}}{(1-\mu)X_{\min}}+1\right)(2X_{\min})^{-1}\,C\,|\delta_0|_1}\right)^2.
\end{aligned}
$$

Furthermore, (5.4) implies that

$$
\sqrt{c_\tau^{\infty}} + (2X_{\min})^{-1}\,c_\tau^{\infty}C\,|\delta_0|_1 < \frac{6X_{\max}^2}{\kappa^2}\lambda s,
$$

which in turn yields that

$$
c_\alpha^{\infty} = \frac{3}{(1-\mu)X_{\min}}\left(\sqrt{c_\tau^{\infty}} + (2X_{\min})^{-1}\,c_\tau^{\infty}C\,|\delta_0|_1\right) < c_\alpha^{*},
$$

and that $c_\tau^{\infty} < c_\tau^{*}$ by construction of c_τ^{*} in $(C.12)$.

Case (ii): Consider the case that

$$
c_\alpha^{(m)} = \frac{6X_{\max}}{(1-\mu)X_{\min}\kappa}\left(c_\alpha^{(m-1)}c_\tau^{(m-1)}C\,|\delta_0|_1\,s\right)^{1/2} =: B_1\sqrt{c_\alpha^{(m-1)}}\sqrt{c_\tau^{(m-1)}}.
$$

where B_1 is defined accordingly. Again, by Lemma 9, we have that

$$\begin{aligned} c_\tau^{(m)} &= c^{-1}\lambda\left((1+\mu)X_{\max}c_\alpha^{(m)} + \sqrt{c_\tau^{(m-1)}} + (2X_{\min})^{-1}c_\tau^{(m-1)}C\left|\delta_0\right|_1\right) \\ &= \left(\frac{\lambda(1+\mu)6X_{\max}^2(C\left|\delta_0\right|_1 s)^{1/2}}{c(1-\mu)X_{\min}\kappa}\sqrt{c_\alpha^{(m-1)}} + \frac{\lambda}{c}\right)\sqrt{c_\tau^{(m-1)}} + \frac{\lambda C\left|\delta_0\right|_1}{c2X_{\min}}c_\tau^{(m-1)} \\ &=: \left(B_2\sqrt{c_\alpha^{(m-1)}} + B_3\right)\sqrt{c_\tau^{(m-1)}} + B_4 c_\tau^{(m-1)}, \end{aligned}$$

by defining B_2, B_3, and B_4 accordingly. As above this system has one fixed point

$$\begin{aligned} c_\tau^{\infty} &= \left(\frac{B_3}{1-B_1B_2-B_4}\right)^2 \\ &= \left(\frac{\lambda/c}{1-\left(\frac{(1+\mu)72X_{\max}^3}{(1-\mu)^2X_{\min}\kappa^2}s+1\right)\frac{C|\delta_0|_1}{c2X_{\min}}\lambda}\right)^2 \end{aligned}$$

and

$$c_\alpha^{\infty} = B_1^2 c_\tau^{\infty} = \left(\frac{6X_{\max}}{(1-\mu)X_{\min}\kappa}\right)^2 C\left|\delta_0\right|_1 s c_\tau^{\infty},$$

provided that $B_1B_2 + B_4 < 1$, which is true under under (5.5). Furthermore, the fixed points c_α^{∞} and c_τ^{∞} of this system is strictly smaller than c_α^* and c_τ^*, respectively, under (5.6) and (5.7) .

Since we have shown that $c_\tau^{\infty} < c_\tau^*$ and $c_\alpha^{\infty} < c_\alpha^*$ in both cases and $c_\tau^{(m)}$ and $c_\alpha^{(m)}$ are strictly decreasing as m increases, the bound in the main theorem is reached within a finite number, say m^*, of iterative applications of Lemma 8 and 9. Therefore, for each case, we have shown that $|\widehat{\alpha}-\alpha_0|_1 \le c_\alpha^*$ and $|\hat{\tau}-\tau_0| \le c_\tau^*$. The bound for the prediction risk can be obtained similarly, and then the bound for the sparsity of the Lasso estimator follows from Lemma 11. Finally, each application of Lemmas 8 and 9 in the chaining argument requires conditioning on $\mathbb{C}(\eta_j)$, $j = 1, ..., m^*$. □

Table 1. List of Variables

Variable Names	Description
Dependent Variable	
gr	Annualized GDP growth rate in the period of 1960–85
Threshold Variables	
gdp60	Real GDP per capita in 1960 (1985 price)
lr	Adult literacy rate in 1960
Covariates	
lgdp60	Log GDP per capita in 1960 (1985 price)
lr	Adult literacy rate in 1960 (only included when $Q = lr$)
ls_k	Log(Investment/Output) annualized over 1960-85; a proxy for the log physical savings rate
lgr_{pop}	Log population growth rate annualized over 1960–85
pyrm60	Log average years of primary schooling in the male population in 1960
pyrf60	Log average years of primary schooling in the female population in 1960
syrm60	Log average years of secondary schooling in the male population in 1960
syrf60	Log average years of secondary schooling in the female population in 1960
hyrm60	Log average years of higher schooling in the male population in 1960
hyrf60	Log average years of higher schooling in the female population in 1960
nom60	Percentage of no schooling in the male population in 1960
nof60	Percentage of no schooling in the female population in 1960
prim60	Percentage of primary schooling attained in the male population in 1960
prif60	Percentage of primary schooling attained in the female population in 1960
pricm60	Percentage of primary schooling complete in the male population in 1960
pricf60	Percentage of primary schooling complete in the female population in 1960
secm60	Percentage of secondary schooling attained in the male population in 1960
secf60	Percentage of secondary schooling attained in the female population in 1960
seccm60	Percentage of secondary schooling complete in the male population in 1960
seccf60	Percentage of secondary schooling complete in the female population in 1960
llife	Log of life expectancy at age 0 averaged over 1960–1985
lfert	Log of fertility rate (children per woman) averaged over 1960–1985
edu/gdp	Government expenditure on eduction per GDP averaged over 1960–85
gcon/gdp	Government consumption expenditure net of defence and education per GDP averaged over 1960–85
revol	The number of revolutions per year over 1960–84
revcoup	The number of revolutions and coups per year over 1960–84
wardum	Dummy for countries that participated in at least one external war over 1960–84
wartime	The fraction of time over 1960-85 involved in external war
lbmp	Log(1+black market premium averaged over 1960–85)
tot	The term of trade shock
lgdp60 × *'educ'*	Product of two covariates (interaction of *lgdp60* and education variables from *pyrm60* to *seccf60*); total 16 variables

Table 2. Model Selection and Estimation Results with $Q = gdp60$

	Linear Model	Threshold Model $\widehat{\tau} = 2898$	
		$\widehat{\beta}$	$\widehat{\delta}$
const.	0.0232	0.0232	-
lgdp60	-0.0153	-0.0120	-
ls_k	0.0033	0.0038	-
lgr_{pop}	0.0018	-	-
pyrf60	0.0027	-	-
syrm60	0.0157	-	-
hyrm60	0.0122	0.0130	-
hyrf60	-0.0389	-	-0.0807
nom60	-	-	2.64×10^{-5}
prim60	-0.0004	-0.0001	-
pricm60	0.0006	-1.73×10^{-4}	-0.35×10^{-4}
pricf60	-0.0006	-	-
secf60	0.0005	-	-
seccm60	0.0010	-	0.0014
llife	0.0697	0.0523	-
lfert	-0.0136	-0.0047	-
edu/gdp	-0.0189	-	-
gcon/gdp	-0.0671	-0.0542	-
revol	-0.0588	-	-
revcoup	0.0433	-	-
wardum	-0.0043	-	-0.0022
wartime	-0.0019	-0.0143	-0.0023
lbmp	-0.0185	-0.0174	-0.0015
tot	0.0971	-	0.0974
lgdp60 × *pyrf60*	-	-3.81×10^{-6}	-
lgdp60 × *syrm60*	-	-	0.0002
lgdp60 × *hyrm60*	-	-	0.0050
lgdp60 × *hyrf60*	-	-0.0003	-
lgdp60 × *nom60*	-	-	8.26×10^{-6}
lgdp60 × *prim60*	-6.02×10^{-7}	-	-
lgdp60 × *prif60*	-3.47×10^{-6}	-	-8.11×10^{-6}
lgdp60 × *pricf60*	-8.46×10^{-6}	-	-
lgdp60 × *secm60*	-0.0001	-	-
lgdp60 × *seccf60*	-0.0002	-2.87×10^{-6}	-
λ	0.0004	0.0034	
$\mathcal{M}(\widehat{\alpha})$	28	26	
# of covariates	46	92	
# of obsesrvations	80	80	

Note: The regularization parameter λ is chosen by the 'leave-one-out' least squares cross validation method. $\mathcal{M}(\widehat{\alpha})$ denotes the number of covariates to be selected by LASSO, and '-' indicates that the regressor is not selected. Recall that $\widehat{\beta}$ is the coefficient when $Q \geq \widehat{\gamma}$ and that $\widehat{\delta}$ is the change of the coefficient value when $Q < \widehat{\gamma}$.

TABLE 3. Model Selection and Estimation Results with $Q = lr$

	Linear Model	Threshold Model $\widehat{\tau} = 82$	
		$\widehat{\beta}$	$\widehat{\delta}$
const.	0.0224	0.0224	-
lgdp60	-0.0159	-0.0099	-
ls_k	0.0038	0.0046	-
syrm60	0.0069	-	-
hyrm60	0.0188	0.0101	-
prim60	-0.0001	-0.0001	-
pricm60	0.0002	0.0001	0.0001
seccm60	0.0004	-	0.0018
llife	0.0674	0.0335	-
lfert	-0.0098	-0.0069	-
edu/gdp	-0.0547	-	-
gcon/gdp	-0.0588	-0.0593	-
revol	-0.0299	-	-
revcoup	0.0215	-	-
wardum	-0.0017	-	-
wartime	-0.0090	-0.0231	-
lbmp	-0.0161	-0.0142	-
tot	0.1333	0.0846	-
lgdp60 × *hyrf60*	-0.0014	-	-0.0053
lgdp60 × *nof60*	1.49×10^{-5}	-	-
lgdp60 × *prif60*	-1.06×10^{-5}	-	-2.66×10^{-6}
lgdp60 × *seccf60*	-0.0001	-	-
λ	0.0011	0.0044	
$\mathcal{M}(\widehat{\alpha})$	22	16	
# of covariates	47	94	
# of observations	70	70	

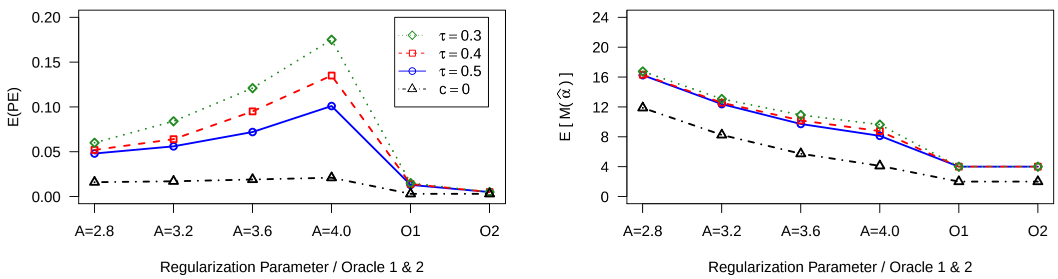
Note: The regularization parameter λ is chosen by the 'leave-one-out' least squares cross validation method. $\mathcal{M}(\widehat{\alpha})$ denotes the number of covariates to be selected by LASSO, and '-' indicates that the regressor is not selected. Recall that $\widehat{\beta}$ is the coefficient when $Q \geq \widehat{\gamma}$ and that $\widehat{\delta}$ is the change of the coefficient value when $Q < \widehat{\gamma}$.

Table 4. Simulation Results with $M = 50$

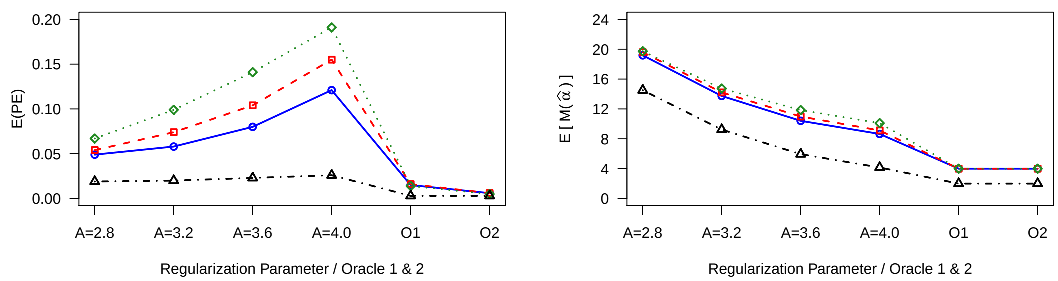
Threshold Parameter	Estimation Method	Constant for λ	Prediction Error (PE) Mean	Median	SD	$E[\mathcal{M}(\widehat{\alpha})]$	$E\|\widehat{\alpha}-\alpha_0\|_1$	$E\|\widehat{\tau}-\tau_0\|_1$
			Jump Scale: $c = 1$					
$\tau_0 = 0.5$	Least Squares	None	0.285	0.276	0.074	100.00	7.066	0.008
	Lasso	$A = 2.8$	0.041	0.030	0.035	12.94	0.466	0.010
		$A = 3.2$	0.048	0.033	0.049	10.14	0.438	0.013
		$A = 3.6$	0.067	0.037	0.086	8.44	0.457	0.024
		$A = 4.0$	0.095	0.050	0.120	7.34	0.508	0.040
	Oracle 1	None	0.013	0.006	0.019	4.00	0.164	0.004
	Oracle 2	None	0.005	0.004	0.004	4.00	0.163	0.000
$\tau_0 = 0.4$	Least Squares	None	0.317	0.304	0.095	100.00	7.011	0.008
	Lasso	$A = 2.8$	0.052	0.034	0.063	13.15	0.509	0.016
		$A = 3.2$	0.063	0.037	0.083	10.42	0.489	0.023
		$A = 3.6$	0.090	0.045	0.121	8.70	0.535	0.042
		$A = 4.0$	0.133	0.061	0.162	7.68	0.634	0.078
	Oracle 1	None	0.014	0.006	0.022	4.00	0.163	0.004
	Oracle 2	None	0.005	0.004	0.004	4.00	0.163	0.000
$\tau_0 = 0.3$	Least Squares	None	2.559	0.511	16.292	100.00	12.172	0.012
	Lasso	$A = 2.8$	0.062	0.035	0.091	13.45	0.602	0.030
		$A = 3.2$	0.089	0.041	0.125	10.85	0.633	0.056
		$A = 3.6$	0.127	0.054	0.159	9.33	0.743	0.099
		$A = 4.0$	0.185	0.082	0.185	8.43	0.919	0.168
	Oracle 1	None	0.012	0.006	0.017	4.00	0.177	0.004
	Oracle 2	None	0.005	0.004	0.004	4.00	0.176	0.000
			Jump Scale: $c = 0$					
N/A	Least Squares	None	6.332	0.460	41.301	100.00	20.936	N/A
	Lasso	$A = 2.8$	0.013	0.011	0.007	9.30	0.266	
		$A = 3.2$	0.014	0.012	0.008	6.71	0.227	
		$A = 3.6$	0.015	0.014	0.009	4.95	0.211	
		$A = 4.0$	0.017	0.016	0.010	3.76	0.204	
	Oracle 1 & 2	None	0.002	0.002	0.003	2.00	0.054	

Note: M denotes the column size of X_i and τ denotes the threshold parameter. Oracle 1 & 2 are estimated by the least squares when sparsity is known and when sparsity and τ_0 are known, respectively. All simulations are based on 400 replications of a sample with 200 observations.

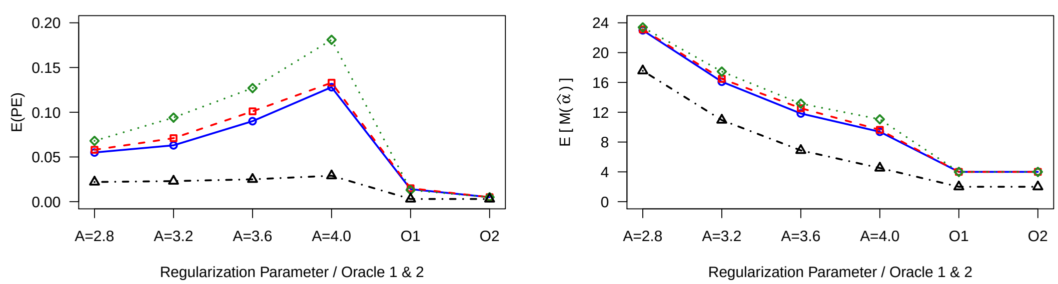
FIGURE 1. Mean Prediction Errors and Mean $\mathcal{M}(\widehat{\alpha})$



$M = 100$

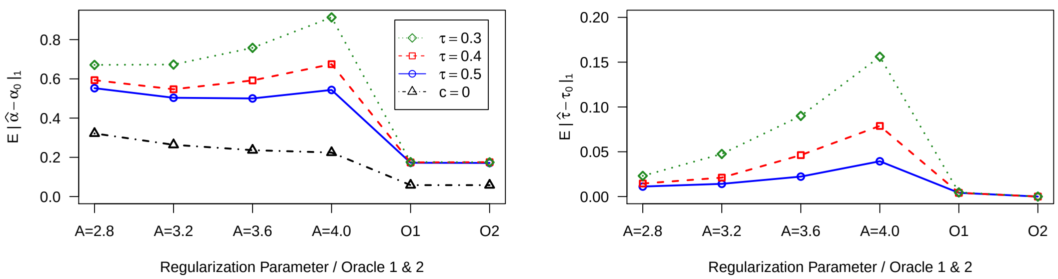


$M = 200$

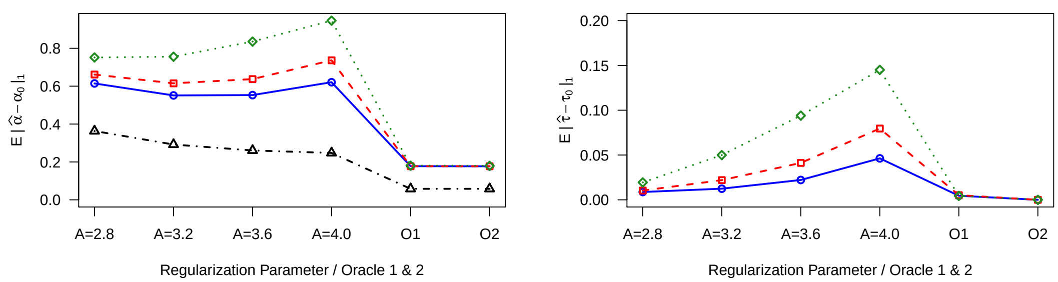


$M = 400$

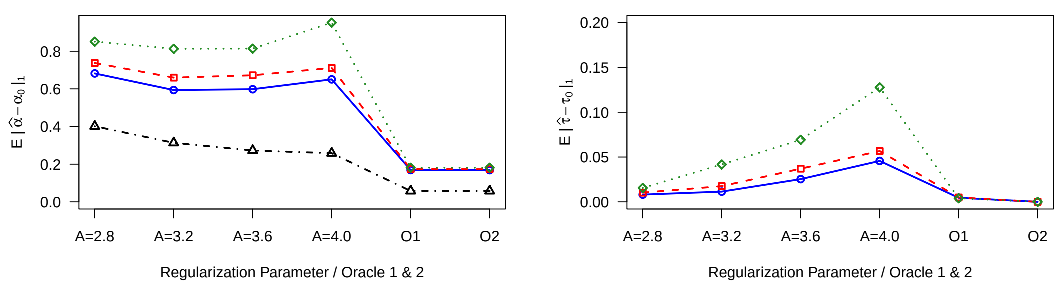
FIGURE 2. Mean ℓ_1-Errors for α and τ



$M = 100$



$M = 200$



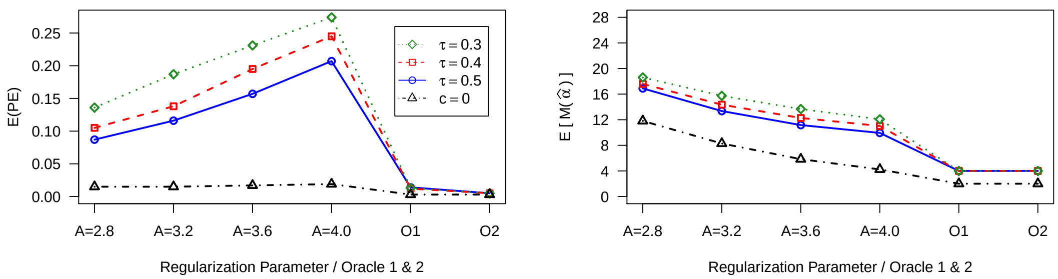
$M = 400$

TABLE 5. Simulation Results with $M = 50$ and $\rho = 0.3$

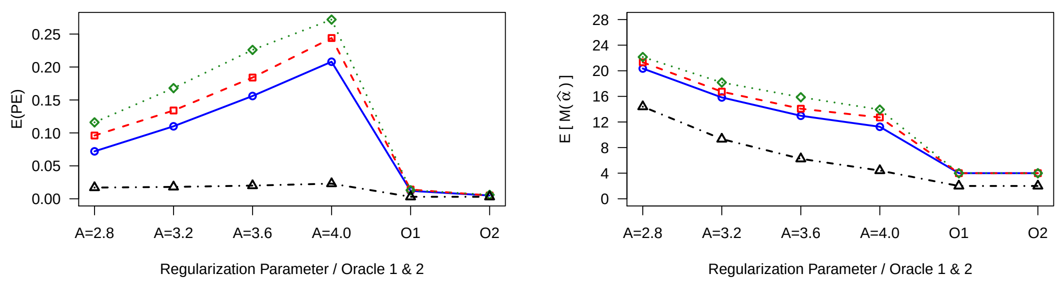
Threshold Parameter	Estimation Method	Constant for λ	Prediction Error (PE) Mean	Median	SD	$E[\mathcal{M}(\widehat{\alpha})]$	$E\lvert\widehat{\alpha}-\alpha_0\rvert_1$	$E\lvert\widehat{\tau}-\tau_0\rvert_1$
			Jump Scale: $c = 1$					
$\tau_0 = 0.5$	Least Squares	None	0.283	0.273	0.075	100.00	7.718	0.010
	Lasso	$A = 2.8$	0.075	0.043	0.087	12.99	0.650	0.041
		$A = 3.2$	0.108	0.059	0.115	10.98	0.737	0.071
		$A = 3.6$	0.160	0.099	0.137	9.74	0.913	0.119
		$A = 4.0$	0.208	0.181	0.143	8.72	1.084	0.166
	Oracle 1	None	0.013	0.006	0.017	4.00	0.169	0.005
	Oracle 2	None	0.005	0.004	0.004	4.00	0.163	0.000
$\tau_0 = 0.4$	Least Squares	None	0.317	0.297	0.099	100.00	7.696	0.010
	Lasso	$A = 2.8$	0.118	0.063	0.123	13.89	0.855	0.094
		$A = 3.2$	0.155	0.090	0.139	11.69	0.962	0.138
		$A = 3.6$	0.207	0.201	0.143	10.47	1.150	0.204
		$A = 4.0$	0.258	0.301	0.138	9.64	1.333	0.266
	Oracle 1	None	0.013	0.007	0.016	4.00	0.168	0.006
	Oracle 2	None	0.005	0.004	0.004	4.00	0.163	0.000
$\tau_0 = 0.3$	Least Squares	None	1.639	0.487	7.710	100.00	12.224	0.015
	Lasso	$A = 2.8$	0.149	0.080	0.136	14.65	1.135	0.184
		$A = 3.2$	0.200	0.233	0.138	12.71	1.346	0.272
		$A = 3.6$	0.246	0.284	0.127	11.29	1.548	0.354
		$A = 4.0$	0.277	0.306	0.116	10.02	1.673	0.408
	Oracle 1	None	0.013	0.006	0.017	4.00	0.182	0.005
	Oracle 2	None	0.005	0.004	0.004	4.00	0.176	0.000
			Jump Scale: $c = 0$					
N/A	Least Squares	None	6.939	0.437	42.698	100.00	23.146	N/A
	Lasso	$A = 2.8$	0.012	0.011	0.007	9.02	0.248	
		$A = 3.2$	0.013	0.011	0.008	6.54	0.214	
		$A = 3.6$	0.014	0.013	0.009	5.00	0.196	
		$A = 4.0$	0.016	0.014	0.010	3.83	0.191	
	Oracle 1 & 2	None	0.002	0.002	0.003	2.00	0.054	

Note: M denotes the column size of X_i and τ denotes the threshold parameter. Oracle 1 & 2 are estimated by the least squares when sparsity is known and when sparsity and τ_0 are known, respectively. All simulations are based on 400 replications of a sample with 200 observations.

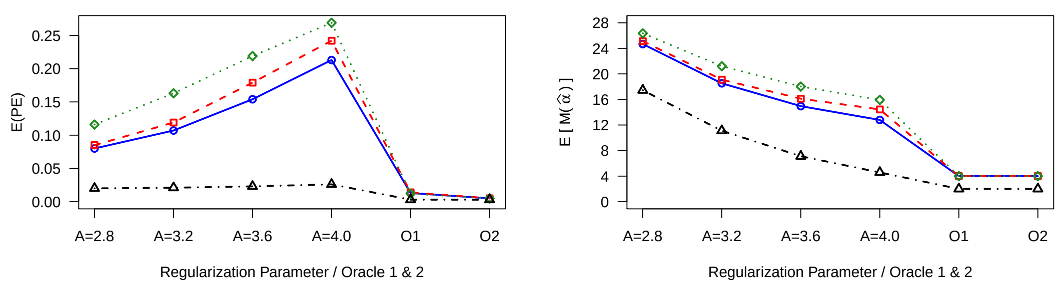
FIGURE 3. Mean Prediction Errors and Mean $\mathcal{M}(\widehat{\alpha})$ when $\rho = 0.3$



$M = 100$

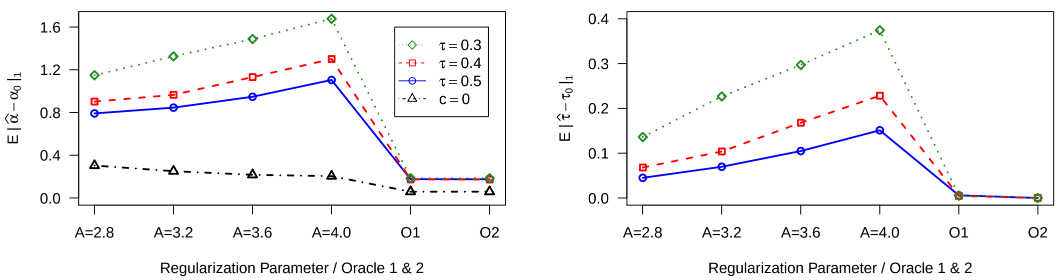


$M = 200$

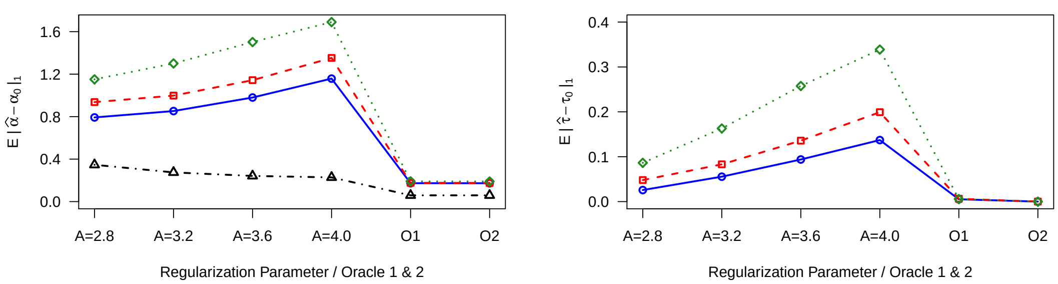


$M = 400$

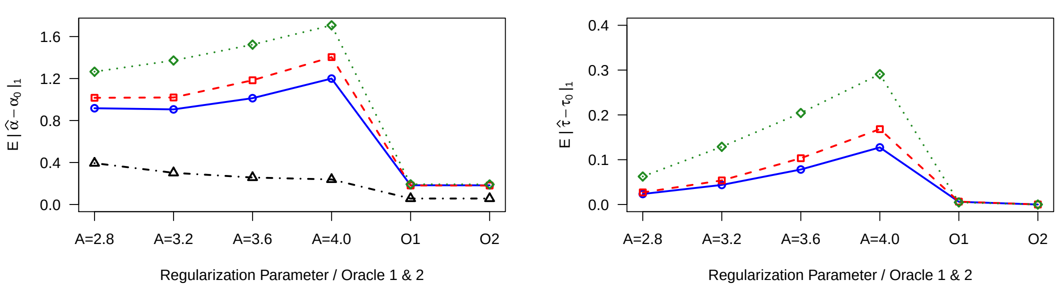
FIGURE 4. Mean ℓ_1-Errors for α and τ when $\rho = 0.3$



$M = 100$

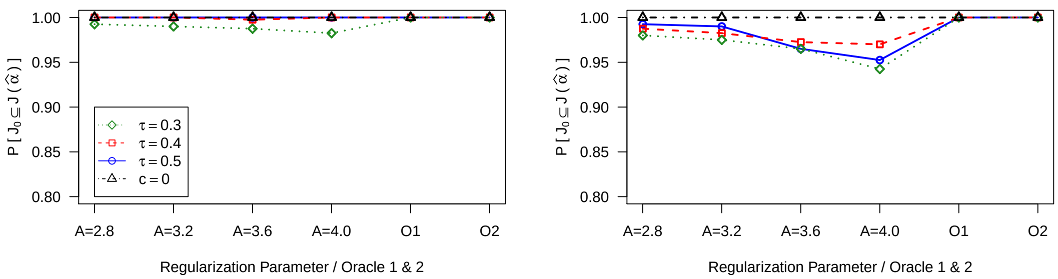


$M = 200$

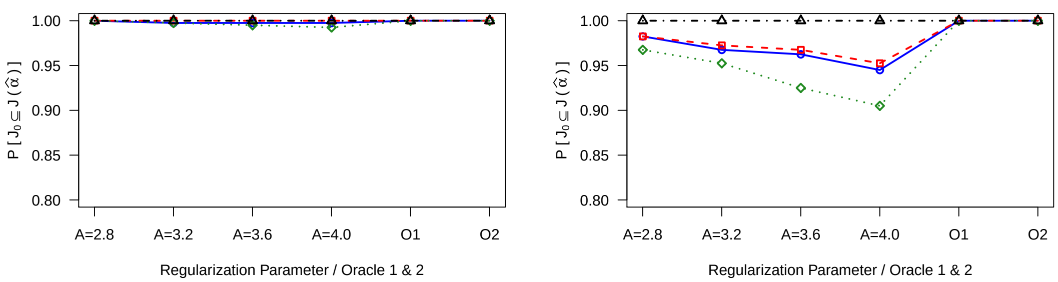


$M = 400$

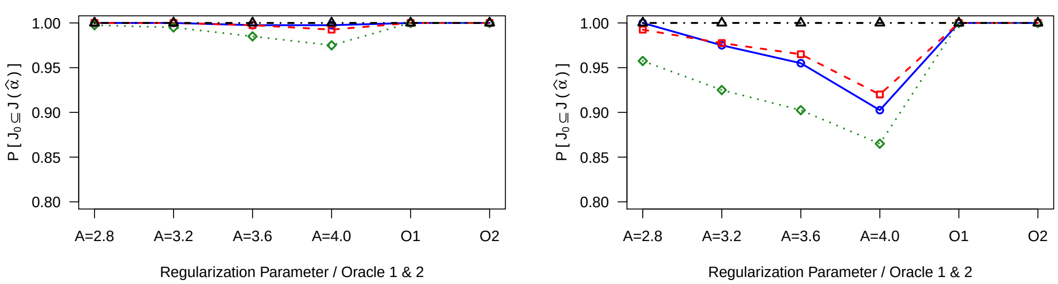
FIGURE 5. Probability of Selecting True Parameters when $\rho = 0$ and $\rho = 0.3$



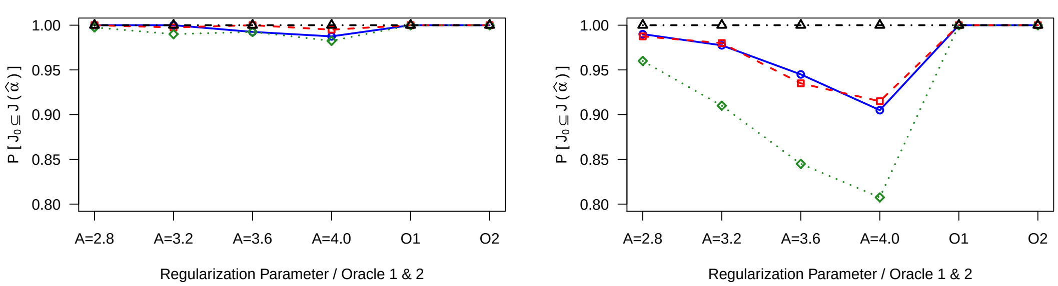
$M = 50$



$M = 100$



$M = 200$



$M = 400$

References

Barro, R. and J. Lee (1994). *Data set for a panel of 139 countries.* Available at http://admin.nber.org/pub/barro.lee/.

Barro, R. and X. Sala-i-Martin (1995). *Economic Growth.* McGraw-Hill. New York.

Belloni, A. and V. Chernozhukov (2011a). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist. 39*(1), 82–130.

Belloni, A. and V. Chernozhukov (2011b). High dimensional sparse econometric models: An introduction. In P. Alquier, E. Gautier, and G. Stoltz (Eds.), *Inverse Problems and High-Dimensional Estimation*, Volume 203 of *Lecture Notes in Statistics*, pp. 121–156. Springer Berlin Heidelberg.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Bradic, J., J. Fan, and J. Jiang (2012). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics 39*(6), 3092–3120.

Bradic, J., J. Fan, and W. Wang (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(3), 325–349.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications.* New York: Springer.

Bunea, F., A. Tsybakov, and M. Wegkamp (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics 1*, 169–194.

Candès, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist. 35*(6), 2313–2351.

Card, D., A. Mas, and J. Rothstein (2008). Tipping and the dynamics of segregation. *Quarterly Journal of Economics 123*(1), 177–218.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of Statistics 21*, 520–533.

Ciuperca, G. (2012). Model selection by LASSO methods in a change-point model. Working Paper arXiv:1107.0865v2. available at http://arxiv.org/abs/1107.0865v2.

Durlauf, S., P. Johnson, and J. Temple (2005). Growth econometrics. *Handbook of economic growth 1*, 555–677.

Durlauf, S. N. and P. A. Johnson (1995). Multiple regimes and cross-country growth behavior. *Journal of Applied Econometrics 10*(4), 365–384.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties,. *Journal of the American Statistical Association 96*, 1348.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*, 101–148.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on 57*(8), 5467–5484.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist. 32*(3), 928–961.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*(3), 575–603.

Harchaoui, Z. and C. Levy-Leduc (2008). Catching change-points with Lasso. In *Advances in Neural Information Processing Systems*, Volume Vol. 20, Cambridge, MA. MIT Press.

Harchaoui, Z. and C. Levy-Leduc (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association 105*(492), 1480–1493.

Huang, J., J. L. Horowitz, and M. S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist. 36*(2), 587–613.

Huang, J., S. G. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models,. *Statistica Sinica 18*, 1603.

Kim, Y., H. Choi, and H.-S. Oh (2008). Smoothly clipped absolute deviation on high dimensions,. *Journal of the American Statistical Association 103*, 1665.

Lee, S., M. Seo, and Y. Shin (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association 106*(493), 220–231.

Lin, W. and J. Lv (2012). High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, forthcoming.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist. 37*(1), 246–270.

Pesaran, M. H. and A. Pick (2007). Econometric issues in the analysis of contagion. *Journal of Economic Dynamics and Control 31*(4), 1245–1277.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. Roy. Statist. Soc. Ser. B 73*(3), 273–282.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach.* New York: Oxford University Press.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics 36*(2), 614–645.

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat. 3*, 1360–1392.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process*. Springer, New York.

Wang, L., Y. Wu, and R. Li (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association 107*(497), 214–222.

Wu, Y. (2008). Simultaneous change point analysis and variable selection in a regression problem. *Journal of Multivariate Analysis 99*(9), 2154 – 2171.

Zhang, N. R. and D. O. Siegmund (2007). Model selection for high dimensional multi-sequence change-point problems. Statistica Sinica Preprint No. SS-10-257. available at http://www3.stat.sinica.edu.tw/preprint/SS-10-257_Preprint.pdf.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*(476), 1418–1429.

Department of Economics, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul, 151-742, Republic of Korea, and The Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK.

E-mail address: sokbae@gmail.com

URL: https://sites.google.com/site/sokbae/.

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: m.seo@lse.ac.uk

URL: http://personal.lse.ac.uk/SEO.

Department of Economics, University of Western Ontario, 1151 Richmond Street N, London, ON N6A 5C2, Canada.

E-mail address: yshin29@uwo.ca

URL: http://publish.uwo.ca/~yshin29.