Oracle Estimation of a Change Point in High Dimensional Quantile Regression*

Sokbae Lee†, Yuan Liao‡, Myung Hwan Seo§ and Youngki Shin¶

15 November 2016

Abstract

In this paper, we consider a high-dimensional quantile regression model where the sparsity structure may differ between two sub-populations. We develop ℓ_1-penalized estimators of both regression coefficients and the threshold parameter. Our penalized estimators not only select covariates but also discriminate between a model with homogeneous sparsity and a model with a change point. As a result, it is not necessary to know or pretest whether the change point is present, or where it occurs. Our estimator of the change point achieves an oracle property in the sense that its asymptotic distribution is the same as if the unknown active sets of regression coefficients were known. Importantly, we establish this oracle property without a perfect covariate selection, thereby avoiding the need for the minimum level condition on the signals of active covariates. Dealing with high-dimensional quantile regression with an unknown change point calls for a new proof technique since the quantile loss function is non-smooth and furthermore the corresponding objective function is non-convex with respect to the change point. The technique developed in this paper is applicable to a general M-estimation framework with a change point, which may be of independent interest. The proposed methods are then illustrated via Monte Carlo experiments and an application to tipping in the dynamics of racial segregation.

Keywords: Variable selection, high-dimensional M-estimation, sparsity, LASSO, SCAD

†Department of Economics, Columbia University, 1022 International Affairs Building 420 West 118th Street, New York, NY 10027, USA; Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK. Email: sl3841@columbia.edu.

‡Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA. Email: yuan.liao@rutgers.edu.

§Department of Economics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-742, Republic of Korea. Email: myunghseo@snu.ac.kr.

¶Economics Discipline Group, University of Technology Sydney, PO Box 123, Broadway NSW 2007, Australia. Email: yshin12@gmail.com

1 Introduction

In this paper, we consider a high-dimensional quantile regression model where the sparsity structure (e.g., identities and effects of contributing regressors) may differ between two sub-populations, thereby allowing for a possible change point in the model. Let $Y \in \mathbb{R}$ be a response variable, $Q \in \mathbb{R}$ be a scalar random variable that determines a possible change point, and $X \in \mathbb{R}^p$ be a p-dimensional vector of covariates. Here, Q can be a component of X, and p is potentially much larger than the sample size n. Specifically, high-dimensional quantile regression with a change point is modelled as follows:

$$Y = X^T \beta_0 + X^T \delta_0 1\{Q > \tau_0\} + U, \tag{1.1}$$

where $(\beta_0^T, \delta_0^T, \tau_0)$ is a vector of unknown parameters and the regression error U satisfies $\mathbb{P}(U \leq 0 | X, Q) = \gamma$ for some known $\gamma \in (0, 1)$. Unlike mean regression, quantile regression analyzes the effects of active regressors on different parts of the conditional distribution of a response variable. Therefore, it allows the sparsity patterns to differ at different quantiles and also handles heterogeneity due to either heteroskedastic variance or other forms of non-location-scale covariate effects. By taking into account a possible change point in the model, we provide a more realistic picture of the sparsity patterns. For instance, when analyzing high-dimensional gene expression data, the identities of contributing genes may depend on the environmental or demographical variables (e.g., exposed temperature, age or weights).

Our paper is closely related to the literature on models with unknown change points (e.g., Tong (1990), Chan (1993), Hansen (1996, 2000), Pons (2003), Kosorok and Song (2007), Seijo and Sen (2011a,b) and Li and Ling (2012) among many others). Recent papers on change points under high-dimensional setups include Enikeeva and Harchaoui (2013); Chan et al. (2014), Frick et al. (2014), Cho and Fryzlewicz (2015), Chan et al. (2016), Callot et al. (2016), and Lee et al. (2016) among others; however, none of these papers consider a change point in high-dimensional quantile regression. The literature on high-dimensional

quantile regression includes Belloni and Chernozhukov (2011), Bradic et al. (2011), Wang et al. (2012), Wang (2013), and Fan et al. (2014) among others. All the aforementioned papers on quantile regression are under the homogeneous sparsity framework (equivalently, assuming that $\delta_0 = 0$ in (1.1)). Ciuperca (2013) considers penalized estimation of a quantile regression model with breaks, but the corresponding analysis is restricted to the case when p is small.

In this paper, we consider estimating regression coefficients $\alpha_0 \equiv (\beta_0^T, \delta_0^T)^T$ as well as the threshold parameter τ_0 and selecting the contributing regressors based on ℓ_1-penalized estimators. One of the strengths of our proposed procedure is that it does not require to know or pretest whether $\delta_0 = 0$ or not, that is, whether the population's sparsity structure and covariate effects are invariant or not. In other words, we do not need to know whether the threshold τ_0 is present in the model.

For a sparse vector $v \in \mathbb{R}^p$, we denote the active set of v as $J(v) \equiv \{j : v_j \neq 0\}$. One of the main contributions of this paper is that our proposed estimator of τ_0 achieves an *oracle property* in the sense that its asymptotic distribution is the same as if the unknown active sets $J(\beta_0)$ and $J(\delta_0)$ were known. Importantly, we establish this oracle property without assuming a perfect covariate selection, thereby avoiding the need for the minimum level condition on the signals of active covariates.

The proposed estimation method in this paper consists of three main steps: in the first step, we obtain the initial estimators of α_0 and τ_0, whose rates of convergence may be suboptimal; in the second step, we re-estimate τ_0 to obtain an improved estimator of τ_0 that converges at the rate of $O_P(n^{-1})$ and achieves the oracle property mentioned above; in the third step, using the second step estimator of τ_0, we update the estimator of α_0. In particular, we propose two alternative estimators of α_0, depending on the purpose of estimation (prediction vs. variable selection).

The most closely related work is Lee et al. (2016). However, there are several important differences: first, Lee et al. (2016) consider a high-dimensional mean regression model with a

homoskedastic normal error and with deterministic covariates; second, their method consists of one-step least squares estimation with an ℓ_1 penalty; third, they derive non-asymptotic oracle inequalities similar to those in Bickel et al. (2009) but do not provide any distributional result on the estimator of the change point. Compared to Lee et al. (2016), dealing with high-dimensional quantile regression with an unknown change point calls for a new proof technique since the quantile loss function is different from the least squares objective function and is non-smooth. In addition, we allow for heteroskesdastic and non-normal regression errors and stochastic covariates. These changes coupled with the fact that the quantile regression objective function is non-convex with respect to the threshold parameter τ_0 raise new challenges. It requires careful derivation and multiple estimation steps to establish the oracle property for the estimator of τ_0 and also to obtain desirable properties of the estimator of α_0. The technique developed in this paper is applicable to a general M-estimation framework with a change point, which may be of independent interest.

One particular application of (1.1) comes from tipping in the racial segregation in social sciences (see, e.g. Card et al., 2008). The empirical question addressed in Card et al. (2008) is whether and the extent to which the neighborhood's white population decreases substantially when the minority share in the area exceeds a tipping point (or change point). In Section 7, we use the US Census tract dataset constructed by Card et al. (2008) and confirm that the tipping exists in the neighborhoods of Chicago.

The remainder of the paper is organized as follows. Section 2 provides an informal description of our estimation methodology. In Section 3, we derive the consistency of the estimators in terms of the excess risk. Further asymptotic properties of the proposed estimators are given in Sections 4 and 5. In Section 6, we present the results of extensive Monte Carlo experiments. Section 7 illustrates the usefulness of our method by applying it to tipping in the racial segregation. Section 8 concludes and Appendix A describes in detail regarding how to construct the confidence interval for τ_0. In Appendix B, we provide a set of regularity assumptions to derive asymptotic properties of the proposed estimators

in Sections 4 and 5. Online supplements are comprised of 6 appendices for all the proofs as well as additional theoretical and numerical results that are left out for the brevity of the paper.

Notation. Throughout the paper, we use $|v|_q$ for the ℓ_q norm for a vector v with $q = 0, 1, 2$. We use $|v|_\infty$ to denote the sup norm. For two sequences of positive real numbers a_n and b_n, we write $a_n \ll b_n$ and equivalently $b_n \gg a_n$ if $a_n = o(b_n)$. If there exists a positive finite constant c such that $a_n = c \cdot b_n$, then we write $a_n \propto b_n$. Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of a matrix A. We use w.p.a.1 to mean "with probability approaching one." We write $\theta_0 \equiv \beta_0 + \delta_0$. For a $2p$ dimensional vector α, let α_J and α_{J^c} denote its subvectors formed by indices in $J(\alpha_0)$ and $\{1, ..., 2p\} \setminus J(\alpha_0)$, respectively. Likewise, let $X_J(\tau)$ denote the subvector of $X(\tau) \equiv (X^T, X^T 1\{Q > \tau\})^T$ whose indices are in $J(\alpha_0)$. The true parameter vectors β_0, δ_0 and θ_0 (except τ_0) are implicitly indexed by the sample size n, and we allow that the dimensions of $J(\beta_0)$, $J(\delta_0)$ and $J(\theta_0)$ can go to infinity as $n \to \infty$. For simplicity, we omit their dependence on n in our notation. We also use the terms 'change point' and 'threshold' interchangeably throughout the paper.

2 Estimators

2.1 Definitions

In this section, we describe our estimation method. We take the check function approach of Koenker and Bassett (1978). Let $\rho(t_1, t_2) \equiv (t_1 - t_2)(\gamma - 1\{t_1 - t_2 \le 0\})$ denote the loss function for quantile regression. Let $\mathcal{A}$ and $\mathcal{T}$ denote the parameter spaces for $\alpha_0 \equiv (\beta_0^T, \delta_0^T)^T$ and τ_0, respectively. For each $\alpha \equiv (\beta, \delta) \in \mathcal{A}$ and $\tau \in \mathcal{T}$, we write $X^T \beta + X^T \delta 1\{Q > \tau\} = X(\tau)^T \alpha$ with the shorthand notation that $X(\tau) \equiv (X^T, X^T 1\{Q > \tau\})^T$. We suppose that the vector of true parameters is defined as the minimizer of the expected loss:

$$(\alpha_0, \tau_0) = \underset{\alpha \in \mathcal{A}, \tau \in \mathcal{T}}{\operatorname{argmin}} \mathbb{E}\left[\rho(Y, X(\tau)^T \alpha)\right]. \tag{2.1}$$

5

By construction, τ_0 is not unique when $\delta_0 = 0$. However, if $\delta_0 = 0$, then the model reduces to the linear quantile regression model in which β_0 is identifiable under the standard assumptions. In Appendix C.1, we provide sufficient conditions under which α_0 and τ_0 are identified when $\delta_0 \neq 0$.

Suppose we observe independent and identically distributed samples $\{Y_i, X_i, Q_i\}_{i \leq n}$. Let $X_i(\tau)$ and $X_{ij}(\tau)$ denote the i-th realization of $X(\tau)$ and j-th element of $X_i(\tau)$, respectively, $i = 1, \ldots, n$ and $j = 1, \ldots, 2p$, so that $X_{ij}(\tau) \equiv X_{ij}$ if $j \leq p$ and $X_{ij}(\tau) \equiv X_{i,j-p}1\{Q_i > \tau\}$ otherwise. Define

$$R_n(\alpha, \tau) \equiv \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i(\tau)^T \alpha) = \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i^T \beta + X_i^T \delta 1\{Q_i > \tau\}).$$

In addition, let $D_j(\tau) \equiv \{n^{-1} \sum_{i=1}^{n} X_{ij}(\tau)^2\}^{1/2}$, $j = 1, \ldots, 2p$.

We describe the main steps of our ℓ_1-penalized estimation method. For some tuning parameter κ_n, define:

$$\textbf{Step 1: } (\breve{\alpha}, \breve{\tau}) = \text{argmin}_{\alpha \in \mathcal{A}, \tau \in \mathcal{T}} R_n(\alpha, \tau) + \kappa_n \sum_{j=1}^{2p} D_j(\tau)|\alpha_j|. \tag{2.2}$$

This step produces an initial estimator $(\breve{\alpha}, \breve{\tau})$. The tuning parameter κ_n is required to satisfy

$$\kappa_n \propto (\log p)(\log n)\sqrt{\frac{\log p}{n}}. \tag{2.3}$$

Note that we take κ_n that converges to zero at a rate slower than the standard $(\log p/n)^{1/2}$ rate in the literature. This modified rate of κ_n is useful in our context to deal with an unknown τ_0. A data-dependent method of choosing κ_n is discussed in Section 2.3.

Remark 2.1. Define $d_j \equiv (\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2)^{1/2}$ and $d_j(\tau) \equiv (\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 1\{Q_i > \tau\})^{1/2}$. Note that $\sum_{j=1}^{2p} D_j(\tau)|\alpha_j| = \sum_{j=1}^{p} d_j|\beta_j| + \sum_{j=1}^{p} d_j(\tau)|\delta_j|$, so that the weight $D_j(\tau)$ adequately balances the regressors; the weight d_j regarding $|\beta_j|$ does not depend on τ, while the weight $d_j(\tau)$ with respect to $|\delta_j|$ does, which takes into account the effect of the threshold τ on the

parameter change δ.

Remark 2.2. The computational cost in Step 1 is the multiple of grid points to the computational time of estimating the linear quantile model with an ℓ_1 penalty, which is solvable in polynomial time (see e.g. Belloni and Chernozhukov (2011) and Koenker and Mizera (2014) among others).

The main purpose of the first step is to obtain an initial estimator of α_0. The achieved convergence rates of this step might be suboptimal due to the uniform control of the score functions over the space $\mathcal{T}$ of the unknown τ_0.

In the second step, we introduce our improved estimator of the change point τ_0. It does not use a penalty term, while using the first step estimator of α_0. Define:

$$\textbf{Step 2: } \widehat{\tau} = \underset{\tau \in \mathcal{T}}{\text{argmin}}\, R_n(\breve{\alpha}, \tau), \tag{2.4}$$

where $\breve{\alpha}$ is the first step estimator of α_0 in (2.2). In Section 4, we show that when τ_0 is identifiable, $\widehat{\tau}$ is consistent for τ_0 at a rate of n^{-1}. Furthermore, we obtain the limiting distribution of $n(\widehat{\tau} - \tau_0)$, and establish conditions under which its asymptotic distribution is the same as if the true α_0 were known, without a perfect model selection on α_0, nor assuming the minimum signal condition on the nonzero elements of α_0.

In the third step, we update the Lasso estimator of α_0 using a different value of the penalization tuning parameter and the second step estimator of τ_0. In particular, we recommend two different estimators of α_0: one for the prediction and the other for the variable selection, serving for different purposes of practitioners. For two different tuning parameters ω_n and

μ_n whose rates will be specified later by (2.7) and (4.1), define:

Step 3a (for prediction):

$$\widehat{\alpha} = \operatorname{argmin}_{\alpha \in \mathcal{A}} R_n(\alpha, \widehat{\tau}) + \omega_n \sum_{j=1}^{2p} D_j(\widehat{\tau})|\alpha_j|, \tag{2.5}$$

Step 3b (for variable selection):

$$\widetilde{\alpha} = \operatorname{argmin}_{\alpha \in \mathcal{A}} R_n(\alpha, \widehat{\tau}) + \mu_n \sum_{j=1}^{2p} w_j D_j(\widehat{\tau})|\alpha_j|, \tag{2.6}$$

where $\widehat{\tau}$ is the second step estimator of τ_0 in (2.4), and the "signal-adaptive" weight w_j in (2.6), motivated by the local linear approximation of the SCAD penalties (Fan and Li, 2001; Zou and Li, 2008), is calculated based on the Step 3a estimator $\widehat{\alpha}$ from (2.5):

$$w_j \equiv \begin{cases} 1, & |\widehat{\alpha}_j| < \mu_n \\ 0, & |\widehat{\alpha}_j| > a\mu_n \\ \frac{a\mu_n - |\widehat{\alpha}_j|}{\mu_n(a-1)} & \mu_n \leq |\widehat{\alpha}_j| \leq a\mu_n. \end{cases}$$

Here $a > 1$ is some prescribed constant, and $a = 3.7$ is often used in the literature. We take this as our choice of a.

Remark 2.3. For $\widehat{\alpha}$ in (2.5), we set ω_n to converge to zero at a rate of $(\log(p \vee n)/n)^{1/2}$:

$$\omega_n \propto \sqrt{\frac{\log(p \vee n)}{n}}, \tag{2.7}$$

which is a more standard rate compared to κ_n in (2.3)). Therefore, the estimator $\widehat{\alpha}$ converges in probability to α_0 faster than $\breve{\alpha}$. In addition, μ_n in (2.6) is chosen to be slightly larger than ω_n for the purpose of the variable selection. A data-dependent method of choosing ω_n as well as μ_n is discussed in Section 2.3. In Sections 4 and 5, we establish conditions under which $\widehat{\alpha}$ achieves the (minimax) optimal rate of convergence in probability for α_0 regardless

of the identifiability of τ_0.

Remark 2.4. It is well known in linear models without the presence of an unknown τ_0 (see, e.g. Bühlmann and van de Geer (2011)) that the Lasso estimator may not perform well for the purpose of the variable selection. The estimator $\widetilde{\alpha}$ defined in Step 3b uses an entry-adaptive weight w_j that corrects the shrinkage bias, and possesses similar merits of the asymptotic unbiasedness of the SCAD penalty. Therefore, we recommend $\widehat{\alpha}$ for the prediction; while suggesting $\widetilde{\alpha}$ for the variable selection.

Remark 2.5. Note that the objective function is non-convex with respect to τ in the first and second steps. However, the proposed estimators can be calculated efficiently using existing algorithms, and we describe the computation algorithms in Section 2.3.

Remark 2.6. Step 2 can be repeated using the updated estimator of α_0 in Step 3. Analogously, Step 3 can be iterated after that. This would give asymptotically equivalent estimators but might improve the finite-sample performance especially when p is very large. Repeating Step 2 might be useful especially when $\breve{\delta} = 0$ in the first step. In this case, there is no unique $\widehat{\tau}$ in Step 2. So, we skip the second step by setting $\widehat{\tau} = \breve{\tau}$ and move to the third step directly. If a preferred estimator of δ_0 in the third step (either $\widehat{\delta}$ or $\widetilde{\delta}$), depending on the estimation purpose, is different from zero, we could go back to Step 2 and re-estimate τ_0. If the third step estimator of δ_0 is also zero, then we conclude that there is no change point and disregard the first-step estimator $\breve{\tau}$ since τ_0 is not identifiable in this case.

2.2 Comparison of Estimators in Step 3

Step 3 defines two estimators for α_0. In this subsection we briefly explain their major differences and purposes. Step 3b is particularly useful when the variable selection consistency is the main objective, yet it often requires the minimum signal condition ($\min_{\alpha_{0j} \neq 0} |\alpha_{0j}|$ is well separated from zero). In contrast, Step 3a does not require the minimum signal condition, and is recommended for prediction purposes. More specifically:

1. If the minimum signal condition (5.1) indeed holds, a perfect variable selection (variable selection consistency) is possible. Indeed, thanks to the signal-adaptive weights, the estimator of Step 3b introduces little shrinkage biases. As a result, we show in Theorem 4.5 that under very mild conditions, this estimator achieves the variable selection consistency. In contrast, Step 3a does not use signal-adaptive weights. In order to achieve the variable selection consistency, it has to rely on much stronger conditions on the design matrix (i.e., the *irrepresentable condition* of Zhao and Yu (2006)) so as to "balance out" the effects of shrinkage biases, and is less adaptive to correlated designs.

2. In the presence of the minimum signal condition, not only does Step 3b achieve the variable selection consistency, it also has a better rate of convergence than Step 3a (Theorem 4.5). The faster rate of convergence is built on the variable selection consistency, and is still a consequence of the signal-adaptive weights. Intuitively, nonzero elements of α_0 are easier to identify and estimate when the signal is strong. Such a phenomenon has been observed in the literature; see, e.g., Fan and Lv (2011) and many papers on variable selections using "folded-concave" penalizations.

3. In the absence of the minimum signal condition, neither method can achieve variable selection consistency. However, it is not a requirement for the prediction purpose. In this case, we recommend the estimator of Step 3a, because it achieves a fast (minimax) rate of convergence (Theorem 4.4), which is useful for predictions.

4. Finally, we show in Theorem 4.7 that without the minimum signal condition, Step 3b, with the signal-adaptive weights, does not perform badly, in the sense that it still results in estimation and prediction consistency. However, the rate of convergence is slower than that of Step 3a.

2.3 Tuning parameter selection

In this subsection, we provide details on how to choose tuning parameters in applications. Recall that our procedure involves three tuning parameters in the penalization: (1) κ_n in Step 1 ought to dominate the score function uniformly over the range of τ, and hence should be slightly larger than the others; (2) ω_n is used in Step 3a for the prediction, and (3) μ_n in Step 3b for the variable selection should be larger than ω_n. Note that the tuning parameters in both Steps 3a and 3b are similar to those of the existing literature since the change point $\widehat{\tau}$ has been estimated.

We build on the data-dependent selection method in Belloni and Chernozhukov (2011). Define

$$\Lambda(\tau) := \max_{1 \leq j \leq 2p} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{X_{ij}(\tau)\left(\gamma - 1\{U_i \leq \gamma\}\right)}{D_j(\tau)} \right|, \tag{2.8}$$

where U_i is simulated from the *i.i.d.* uniform distribution on the interval $[0, 1]$; γ is the quantile of interest (e.g. $\gamma = 0.5$ for median regression). Note that $\Lambda(\tau)$ is a stochastic process indexed by τ. Let $\overline{\Lambda}_{1-\epsilon^*}$ be the $(1 - \epsilon^*)$-quantile of $\sup_{\tau \in \mathcal{T}} \Lambda(\tau)$, where ϵ^* is a small positive constant that will be selected by a user. Then, we select the tuning parameter in Step 1 by $\kappa_n = c_1 \cdot \overline{\Lambda}_{1-\epsilon^*}$. Similarly, let $\Lambda_{1-\epsilon^*}(\widehat{\tau})$ be the $(1 - \epsilon^*)$-quantile of $\Lambda(\widehat{\tau})$, where $\widehat{\tau}$ is chosen in Step 2. We select ω_n and μ_n in Step 3 by $\omega_n = c_1 \cdot \Lambda_{1-\epsilon^*}(\widehat{\tau})$ and $\mu_n = c_2 \cdot \omega_n$. It is also necessary to choose $\mathcal{T}$ in applications. In our Monte Carlo experiments in Section 6, we take $\mathcal{T}$ to be the interval from the 15th percentile to the 85th percentile of the empirical distribution of the threshold variable Q_i. For example, Hansen (1996) employed the same range in his application to U.S. GNP dynamics.

Based on the suggestions of Belloni and Chernozhukov (2011) and some preliminary simulations, we choose to set $c_1 = 1.1$, $c_2 = \log \log n$, and $\epsilon^* = 0.1$. In addition, recall that we set $a = 3.7$ when calculating the SCAD weights w_j in Step 3b following the convention in the literature (e.g. Fan and Li (2001) and Loh and Wainwright (2013)). In Step 1, we

first solve the lasso problem for α given each grid point of $\tau \in \mathcal{T}$. Then, we choose $\breve{\tau}$ and the corresponding $\breve{\alpha}(\breve{\tau})$ that minimize the objective function. Step 2 can be solved simply by the grid search. Step 3 is a standard lasso quantile regression estimation given $\widehat{\tau}$, whose numerical implementation is well established. We use the `rq()` function of the R 'quantreg' package with the `method = "lasso"` in each implementation of the standard lasso quantile regression estimation (Koenker, 2016).

3   Risk Consistency

Given the loss function $\rho(t_1, t_2) \equiv (t_1 - t_2)(\gamma - 1\{t_1 - t_2 \leq 0\})$ for the quantile regression model, define the *excess risk* to be

$$R(\alpha, \tau) \equiv \mathbb{E}\rho(Y, X(\tau)^T \alpha) - \mathbb{E}\rho(Y, X(\tau_0)^T \alpha_0). \tag{3.1}$$

By the definition of (α_0, τ_0) in (2.1), we have that $R(\alpha, \tau) \geq 0$ for any $\alpha \in \mathcal{A}$ and $\tau \in \mathcal{T}$. What we mean by the "risk consistency" here is that the excess risk converges in probability to zero for the proposed estimators. The other asymptotic properties of the proposed estimators will be presented in Sections 4 and 5.

In this section, we begin by stating regularity conditions that are needed to develop our first theoretical result. Recall that X_{ij} denotes the jth element of X_i.

Assumption 1 (Setting).   *(i) The data $\{(Y_i, X_i, Q_i)\}_{i=1}^n$ are independent and identically distributed. Furthermore, for all j and every integer $m \geq 1$, there is a constant $K_1 < \infty$ such that $\mathbb{E}\,|X_{ij}|^m \leq \frac{m!}{2} K_1^{m-2}$.*

(ii) $\mathbb{P}(\tau_1 < Q \leq \tau_2) \leq K_2(\tau_2 - \tau_1)$ for any $\tau_1 < \tau_2$ and some constant $K_2 < \infty$.

(iii) $\alpha_0 \in \mathcal{A} \equiv \{\alpha : |\alpha|_\infty \leq M_1\}$ for some constant $M_1 < \infty$, and $\tau_0 \in \mathcal{T} \equiv [\underline{\tau}, \overline{\tau}]$. Further-

more, the probability of $\{Q < \underline{\tau}\}$ and that of $\{Q > \overline{\tau}\}$ are strictly positive, and

$$\sup_{j \leq p} \sup_{\tau \in \mathcal{T}} \mathbb{E}[X_{ij}^2 | Q = \tau] < \infty.$$

(iv) There exist universal constants $\underline{D} > 0$ and $\overline{D} > 0$ such that w.p.a.1,

$$0 < \underline{D} \leq \min_{j \leq 2p} \inf_{\tau \in \mathcal{T}} D_j(\tau) \leq \max_{j \leq 2p} \sup_{\tau \in \mathcal{T}} D_j(\tau) \leq \overline{D} < \infty.$$

(v) $\mathbb{E}\left[\left(X^T \delta_0\right)^2 | Q = \tau\right] \leq M_2 |\delta_0|_2^2$ for all $\tau \in \mathcal{T}$ and for some constant M_2 satisfying $0 < M_2 < \infty$.

In addition to the random sampling assumption, condition (i) imposes mild moment restrictions on X. Condition (ii) imposes a weak restriction that the probability that $Q \in (\tau_1, \tau_2]$ is bounded by a constant times $(\tau_2 - \tau_1)$. Condition (iii) assumes that the parameter space is compact and that the support of Q is strictly larger than $\mathcal{T}$. These conditions are standard in the literature on change-point and threshold models (e.g., Seijo and Sen (2011a,b)). Condition (iii) also assumes that the conditional expectation of $\mathbb{E}[X_{ij}^2 | Q = \cdot]$ is bounded on $\mathcal{T}$ uniformly in j. Condition (iv) requires that each regressor be of the same magnitude uniformly over the threshold τ. As the data-dependent weights $D_j(\tau)$ are the sample second moments of the regressors, it is not stringent to assume them to be bounded away from both zero and infinity. Condition (v) puts some weak upper bound on $\mathbb{E}[\left(X^T \delta_0\right)^2 | Q = \tau]$ for all $\tau \in \mathcal{T}$ when $\delta_0 \neq 0$. A simple sufficient condition for condition (v) is that the eigenvalues of $\mathbb{E}[X_{J(\delta_0)} X_{J(\delta_0)}^T | Q = \tau]$ are bounded uniformly in τ, where $X_{J(\delta_0)}$ denotes the subvector of X corresponding to the nonzero components of δ_0.

Throughout the paper, we let $s \equiv |J(\alpha_0)|_0$, namely the cardinality of $J(\alpha_0)$. We allow that $s \to \infty$ as $n \to \infty$ and will give precise regularity conditions regarding its growth rates. The following theorem is concerned about the convergence of $R(\breve{\alpha}, \breve{\tau})$ with the first step estimator.

Theorem 3.1 (Risk Consistency)**.** *Let Assumption 1 hold. Suppose that the tuning param-eter κ_n satisfies* (2.3). *Then,* $R(\breve{\alpha}, \breve{\tau}) = O_P\left(\kappa_n s\right).$

Note that Theorem 3.1 holds regardless of the identifiability of τ_0 (that is, whether $\delta_0 = 0$ or not). In addition, the rate $O_P(\kappa_n s)$ is achieved regardless of whether $\kappa_n s$ converges, and we have the risk consistency if $\kappa_n s \to 0$ as $n \to \infty$. The restriction on s is slightly stronger than that of the standard result $s = o(\sqrt{n/\log p})$ in the literature for the M-estimation (see, e.g. van de Geer (2008) and Chapter 6.6 of Bühlmann and van de Geer (2011)) since the objective function $\rho(Y, X(\tau)^T \alpha)$ is non-convex in τ, due to the unknown change-point.

Remark 3.1. The extra logarithmic factor $(\log p)(\log n)$ in the definition of κ_n (see (2.3)) is due to the existence of the unknown and possibly non-identifiable threshold parameter τ_0. In fact, an inspection of the proof of Theorem 3.1 reveals that it suffices to assume that κ_n satisfies $\kappa_n \gg \log_2(p/s)[\log(np)/n]^{1/2}$. The term $\log_2(p/s)$ and the additional $(\log n)^{1/2}$ term inside the brackets are needed to establish the stochastic equicontinuity of the empirical process

$$\nu_n\left(\alpha, \tau\right) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[\rho\left(Y_i, X_i\left(\tau\right)^T \alpha\right) - \mathbb{E}\rho\left(Y, X\left(\tau\right)^T \alpha\right) \right]$$

uniformly over $(\alpha, \tau) \in \mathcal{A} \times \mathcal{T}$.

In Appendix C.2, we show that an improved rate of convergence, $O_P\left(\omega_n s\right)$, is possible for the excess risk by taking the second and third steps of estimation.

4   Asymptotic Properties: Case I. $\delta_0 \neq 0$

Sections 4 and 5 provide asymptotic properties of the proposed estimators. In Appendix B, we list a set of assumptions that are needed to derive these properties, in addition to Assumption 1. We first establish the consistency of $\breve{\tau}$ for τ_0.

Theorem 4.1 (Consistency of $\breve{\tau}$)**.** *Let Assumptions 1, 2, 5, and 6 hold. Furthermore, assume that $\kappa_n s = o(1)$. Then,* $\breve{\tau} \xrightarrow{P} \tau_0.$

The following theorem presents the rates of convergence for the first step estimators of α_0 and τ_0. Recall that κ_n is the first-step penalization tuning parameter that satisfies (2.3).

Theorem 4.2 (Rates of Convergence When $\delta_0 \neq 0$). *Suppose that $\kappa_n s^2 \log p = o(1)$. Then under Assumptions 1-6, we have:*

$$|\breve{\alpha} - \alpha_0|_1 = O_P(\kappa_n s), \ R(\breve{\alpha}, \breve{\tau}) = O_P(\kappa_n^2 s), \quad and \quad |\breve{\tau} - \tau_0| = O_P(\kappa_n^2 s).$$

In Theorem 3.1, we have that $R(\breve{\alpha}, \breve{\tau}) = O_P(\kappa_n s)$. The improved rate of convergence for $R(\breve{\alpha}, \breve{\tau})$ in Theorem 4.2 is due to additional assumptions (in particular, compatibility conditions in Assumption 3 among others). It is worth noting that $\breve{\tau}$ converges to τ_0 faster than the standard parametric rate of $n^{-1/2}$, as long as $s^2 (\log p)^6 (\log n)^4 = o(n)$. The main reason for such *super-consistency* is that the objective function behaves locally linearly around τ_0 with a kink at τ_0, unlike in the regular estimation problem where the objective function behaves locally quadratically around the true parameter value. Moreover, the achieved convergence rate for $\breve{\alpha}$ is nearly minimax optimal, with an additional factor $(\log p)(\log n)$ compared to the rate of regular Lasso estimation (e.g., Bickel et al. (2009); Raskutti et al. (2011)). This factor arises due to the unknown change-point τ_0. We will improve the rates of convergence for both τ_0 and α_0 further by taking the second and third steps of estimation.

Recall that the second-step estimator of τ_0 is defined as

$$\widehat{\tau} = \underset{\tau \in \mathcal{T}}{\operatorname{argmin}} \, R_n(\breve{\alpha}, \tau),$$

where $\breve{\alpha}$ is the first step estimator of α_0 in (2.2). Consider an oracle case for which α in $R_n(\alpha, \tau)$ is fixed at α_0. Let $R_n^*(\tau) = R_n(\alpha_0, \tau)$ and

$$\widetilde{\tau} = \underset{\tau \in \mathcal{T}}{\operatorname{argmin}} \, R_n^*(\tau).$$

We now give one of the main results of this paper.

Theorem 4.3 (Oracle Estimation of τ_0). *Let Assumptions 1-6 hold. Furthermore, suppose that $\kappa_n s^2 \log p = o(1)$. Then, we have that*

$$\widehat{\tau} - \widetilde{\tau} = o_P\left(n^{-1}\right).$$

Furthermore, $n\left(\widehat{\tau} - \tau_0\right)$ converges in distribution to the smallest minimizer of a compound Poisson process, which is given by

$$M\left(h\right) \equiv \sum_{i=1}^{N_1(-h)} \rho_{1i} 1\left\{h < 0\right\} + \sum_{i=1}^{N_2(h)} \rho_{2i} 1\left\{h \geq 0\right\},$$

where N_1 and N_2 are Poisson processes with the same jump rate $f_Q\left(\tau_0\right)$, and $\{\rho_{1i}\}$ and $\{\rho_{2i}\}$ are two sequences of independent and identically distributed random variables. The distributions of ρ_{1i} and ρ_{2i}, respectively, are identical to the conditional distributions of $\dot{\rho}\left(U_i - X_i^T \delta_0\right) - \dot{\rho}\left(U_i\right)$ and $\dot{\rho}\left(U_i + X_i^T \delta_0\right) - \dot{\rho}\left(U_i\right)$ given $Q_i = \tau_0$, where $\dot{\rho}\left(t\right) \equiv t\left(\gamma - 1\left\{t \leq 0\right\}\right)$ and $U_i \equiv Y_i - X_i^T \beta_0 - X_i^T \delta_0 1\left\{Q_i > \tau_0\right\}$ for each $i = 1, \ldots, n$. Here, N_1, N_2, $\{\rho_{1i}\}$ and $\{\rho_{2i}\}$ are mutually independent.

The first conclusion of Theorem 4.3 establishes that the second step estimator of τ_0 is an oracle estimator in the sense that it is asymptotically equivalent to the infeasible, oracle estimator $\widetilde{\tau}$. As emphasized in the introduction, the oracle property is obtained without relying on the perfect model selection in the first step nor on the existence of the minimum signal condition on active covariates. The second conclusion of Theorem 4.3 follows from combining well-known weak convergence results in the literature (see e.g. Pons (2003); Kosorok and Song (2007); Lee and Seo (2008)) with the argmax continuous mapping theorem by Seijo and Sen (2011b).

Remark 4.1. Li and Ling (2012) propose a numerical approach for constructing a confidence interval by simulating a compound Poisson process in the context of least squares estimation. We adopt their approach to simulate the compound Poisson process for quantile regression.

See Appendix A for a detailed description of how to construct a confidence interval for τ_0.

We now consider the Step 3a estimator of α_0 defined in (2.5). Recall that ω_n is the Step 3a penalization tuning parameter that satisfies (2.7).

Theorem 4.4 (Improved Rates of Convergence When $\delta_0 \neq 0$). *Suppose that $\kappa_n s^2 \log p = o(1)$. Then under Assumptions 1-6,*

$$|\widehat{\alpha} - \alpha_0|_1 = O_P(\omega_n s) \quad and \quad R(\widehat{\alpha}, \widehat{\tau}) = O_P(\omega_n^2 s).$$

Theorem 4.4 shows that the estimator $\widehat{\alpha}$ defined in Step 3a achieves the optimal rate of convergence in terms of prediction and estimation. In other words, when ω_n is proportional to $\{\log(p \vee n)/n\}^{1/2}$ in equation (2.7) and p is larger than n, it obtains the minimax rates as in e.g., Raskutti et al. (2011).

As we mentioned in Section 2, the Step 3b estimator of α_0 has the purpose of the variable selection. The nonzero components of $\widetilde{\alpha}$ are expected to identify contributing regressors. Partition $\widetilde{\alpha} = (\widetilde{\alpha}_J, \widetilde{\alpha}_{J^c})$ such that $\widetilde{\alpha}_J = (\widetilde{\alpha}_j : j \in J(\alpha_0))$ and $\widetilde{\alpha}_{J^c} = (\widetilde{\alpha}_j : j \notin J(\alpha_0))$. Note that $\widetilde{\alpha}_J$ consists of the estimators of β_{0J} and δ_{0J}, whereas $\widetilde{\alpha}_{J^c}$ consists of the estimators of all the zero components of β_0 and δ_0. Let $\alpha_{0J}^{(j)}$ denote the j-th element of α_{0J}.

We now establish conditions under which the estimator $\widetilde{\alpha}$ defined in Step 3b has the *change-point-oracle properties*, meaning that it achieves the variable selection consistency and has the limiting distributions as though the identities of the important regressors and the location of the change point were known.

Theorem 4.5 (Variable Selection When $\delta_0 \neq 0$). *Suppose that $\kappa_n s^2 \log p = o(1)$, $s^4 \log s = o(n)$, and*

$$\omega_n + s\sqrt{\frac{\log s}{n}} \ll \mu_n \ll \min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|. \tag{4.1}$$

Then under Assumptions 1-6, we have: (i)

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right), \quad |\widetilde{\alpha}_J - \alpha_{0J}|_1 = O_P\left(s\sqrt{\frac{\log s}{n}}\right),$$

(ii)

$$P(\widetilde{\alpha}_{J^c} = 0) \to 1,$$

and (iii)

$$R(\widetilde{\alpha}, \widehat{\tau}) = O_P\left(\mu_n s\sqrt{\frac{\log s}{n}}\right).$$

We see that (4.1) provides a condition on the strength of the signal via $\min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|$, and the tuning parameter in Step 3b should satisfy $\omega_n \ll \mu_n$ and $s^2 \log s/n \ll \mu_n^2$. Hence the variable selection consistency demands a larger tuning parameter than in Step 3a.

To conduct statistical inference, we now discuss the asymptotic distribution of $\widetilde{\alpha}_J$. Define $\widehat{\alpha}_J^* \equiv \operatorname{argmin}_{\alpha_J} R_n^*(\alpha_J, \tau_0)$. Note that the asymptotic distribution for $\widehat{\alpha}_J^*$ corresponds to an oracle case that we know τ_0 as well as the true active set $J(\alpha_0)$ *a priori*. The limiting distribution of $\widetilde{\alpha}_J$ is the same as that of $\widehat{\alpha}_J^*$. Hence, we call this result the *change-point-oracle property* of the Step 3b estimator and the following theorem establishes this property.

Theorem 4.6 (Change-Point-Oracle Properties)**.** *Suppose that all the conditions imposed in Theorem 4.5 are satisfied. Furthermore, assume that $\frac{\partial}{\partial \alpha} E\left[\rho\left(Y, X^T \alpha\right) | Q = t\right]$ exists for all t in a neighborhood of τ_0 and all its elements are continuous and bounded, and that $s^3(\log s)(\log n) = o(n)$. Then, we have that $\widetilde{\alpha}_J = \widehat{\alpha}_J^* + o_P(n^{-1/2})$.*

Since the sparsity index (s) grows at a rate slower than the sample size (n), it is straightforward to establish the asymptotic normality of a linear transformation of $\widetilde{\alpha}_J$, i.e., $\mathbf{L}\widetilde{\alpha}_J$, where $\mathbf{L} : \mathbb{R}^s \to \mathbb{R}$ with $|\mathbf{L}|_2 = 1$, by combing the existing results on quantile regression with parameters of increasing dimension (see, e.g. He and Shao (2000)) with Theorem 4.6.

Remark 4.2. Without the condition on the strength of minimal signals, it may not be possible to achieve the variable selection consistency or establish change-point-oracle properties.

However, the following theorem shows that the SCAD-weighted penalized estimation can still achieve a satisfactory rate of convergence in estimation of α_0 without the condition that $\mu_n \ll \min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|$. Yet, the rates of convergence are slower than those of Theorem 4.5.

Theorem 4.7 (Satisfactory Rates Without Minimum Signal Condition). *Assume that Assumptions 1-6 hold. Suppose that $\kappa_n s^2 \log p = o(1)$ and $\omega_n \ll \mu_n$. Then, without the lower bound requirement on $\min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|$, we have that $|\widetilde{\alpha} - \alpha_0|_1 = O_P(\mu_n s)$. In addition, $R(\widetilde{\alpha}, \widehat{\tau}) = O_P(\mu_n^2 s)$.*

5 Asymptotic Properties: Case II. $\delta_0 = 0$

In this section, we show that our estimators have desirable results even if there is no change point in the true model. The case of $\delta_0 = 0$ corresponds to the high-dimensional linear quantile regression model. Since $X^T \beta_0 + X^T \delta_0 1\{Q > \tau_0\} = X^T \beta_0$, τ_0 is non-identifiable, and there is no structural change on the coefficient. But a new analysis different from that of the standard high-dimensional model is still required because in practice we do not know whether $\delta_0 = 0$ or not. Thus, the proposed estimation method still estimates τ_0 to account for possible structural changes. The following results show that in this case, the first step estimator of α_0 will asymptotically behave as if $\delta_0 = 0$ were *a priori* known.

Theorem 5.1 (Rates of Convergence When $\delta_0 = 0$). *Suppose that $\kappa_n s = o(1)$. Then under Assumptions 1-4, we have that*

$$|\breve{\alpha} - \alpha_0|_1 = O_P(\kappa_n s) \quad and \quad R(\breve{\alpha}, \breve{\tau}) = O_P(\kappa_n^2 s).$$

The results obtained in Theorem 5.1 combined with those obtained in Theorem 4.2 imply that the first step estimatior performs equally well in terms of rates of convergence for both the ℓ_1 loss for $\breve{\alpha}$ and the excess risk regardless of the existence of the threshold effect. It is straightforward to obtain an improved rate result for the Step 3a estimator, equivalent to

Theorem 4.4 under Assumptions 1-4. We omit the details for brevity.

We now give a result that is similar to Theorem 4.5 and Theorem 4.7.

Theorem 5.2 (Variable Selection When $\delta_0 = 0$). *Suppose that* $\kappa_n s = o(1)$, $s^4 \log s = o(n)$, $\omega_n + s\sqrt{\frac{\log s}{n}} \ll \mu_n$, *and Assumptions 1-4 hold. We have:*
(i) If the minimum signal condition holds:

$$\mu_n = o\left(\min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}| \right), \tag{5.1}$$

then

$$\left| \widetilde{\beta}_J - \beta_{0J} \right|_2 = O_P\left(\sqrt{\frac{s \log s}{n}} \right), \quad \left| \widetilde{\beta}_J - \beta_{0J} \right|_1 = O_P\left(s\sqrt{\frac{\log s}{n}} \right),$$

$$P(\widetilde{\beta}_{J^c} = 0) \to 1, \quad P(\widetilde{\delta} = 0) \to 1, \quad and \quad R(\widetilde{\alpha}, \widehat{\tau}) = O_P\left(\mu_n s\sqrt{\frac{\log s}{n}} \right).$$

(ii) Without the minimum signal condition (5.1), we have:

$$R(\widetilde{\alpha}, \widehat{\tau}) = O_P(\mu_n^2 s), \quad |\widetilde{\alpha} - \alpha_0|_1 = O_P(s\mu_n).$$

Theorem 5.2 demonstrates that when there is in fact no change point, our estimator for δ_0 is exactly zero with a high probability. Therefore, the estimator can also be used as a diagnostic tool to check whether there exists any change point. Results similar to Theorems 4.6 can be established straightforwardly as well; however, their details are omitted for brevity.

6 Monte Carlo Experiments

In this section we provide the results of Monte Carlo experiments. The baseline model is based on the following data generating process: for $i = 1, \ldots, n$,

$$Y_i = X_i'(\beta_0 + \xi_{10}U_i) + 1\{Q_i > \tau_0\}X_i'(\delta_0 + \xi_{20}U_i), \tag{6.1}$$

where U_i follows $N(0, 0.5^2)$, and Q_i follows the uniform distribution on the interval $[0,1]$. The p-dimensional covariate X_i is composed of a constant and Z_i, i.e. $X := (1, Z_i^T)^T$, where Z_i follows the multivariate normal distribution $N(0, \Sigma)$ with a covariance matrix $\Sigma_{ij} = (1/2)^{|i-j|}$. Here, the variables U_i, Q_i and Z_i are independent of each other. Note that the conditional γ-quantile of Y_i given (X_i, Q_i) has the form:

$$Quant_\gamma(Y_i|X_i, Q_i) = X_i'\beta_\gamma + 1\{Q_i < \tau_0\}X_i'\delta_\gamma, \tag{6.2}$$

where $\beta_\gamma = \beta_0 + \xi_{10} \cdot Quant_\gamma(U)$ and $\delta_\gamma = \delta_0 + \xi_{20} \cdot Quant_\gamma(U)$.

We consider three quantile regression models with $\gamma = 0.25, 0.5$, and 0.75. The p-dimensional parameters $\beta_0, \delta_0, \xi_{10}$, and ξ_{20} are set to $\beta_0 = (0, Quant_{0.75}(U) \approx 0.34, 0, \ldots, 0)$, $\delta_0 = (0, 1, 0, \ldots, 0)$, $\xi_{10} = (0, 1, 0, \ldots, 0)$, and $\xi_{20} = (0, 0, 0, \ldots, 0)$, respectively. Because of the heteroskedasticity, the true parameter value β_γ at each quantile is $\beta_{0.25} = (0, \ldots, 0)$, $\beta_{0.5} = (0, 0.34, \ldots, 0)$, and $\beta_{0.75} = (0, 0.68, \ldots, 0)$. Note that nonzero coefficients are different between when $\gamma = 0.25$ and when $\gamma = 0.5$ or $\gamma = 0.75$.

We set the change point parameter $\tau_0 = 0.5$ unless it is specified differently. The sample sizes are set to $n = 200$ and 400. The dimension of X_i is set to $p = 250$. Note that we have 500 regressors in total. The change point τ is estimated over grid points of the sample observations $\{Q_i\}$, where the range is limited to those between the 0.15-quantile and the 0.85-quantile. We conduct 1,000 replications of each design.

We compare estimation results of each step. To assess the performance of our estimators, we also compare the results with two "oracle estimators". Specifically, Oracle 1 knows the true active set $J(\alpha_\gamma)$ and the change point parameter τ_0, and Oracle 2 knows only $J(\alpha_\gamma)$. The threshold parameter τ_0 is re-estimated in Steps 3a and 3b using updated estimates of α_γ.

Tables 1–3 summarize the simulation results. We abuse notation slightly and denote all estimators by $(\widehat{\alpha}, \widehat{\tau})$. They would be understood as $(\breve{\alpha}, \breve{\tau})$ in Step 1, $\widehat{\tau}$ in Step 2, and so on.

21

Table 1: Baseline Model: $\gamma = 0.25$

	Excess Risk	E$[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c}$)	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop.
$n = 200$							
Oracle 1	0.004	NA	0.005 (NA / NA)	0.203	NA	NA	NA
Oracle 2	0.013	NA	0.005 (NA / NA)	0.434	0.012	0.925	NA
Step 1	0.032	4.266	0.433 (0.389 / 0.044)	0.727	0.013	0.943	0.032
Step 2	0.032	NA	NA (NA / NA)	0.705	0.012	0.952	NA
Step 3a	0.031	4.249	0.413 (0.370 / 0.043)	0.691	0.012	0.951	0.032
Step 3b	0.022	1.173	0.281 (0.221 / 0.060)	0.556	0.012	0.928	0.686
$n = 400$							
Oracle 1	0.002	NA	0.003 (NA / NA)	0.145	NA	NA	NA
Oracle 2	0.006	NA	0.003 (NA / NA)	0.313	0.005	0.958	NA
Step 1	0.017	4.352	0.214 (0.193 / 0.021)	0.502	0.006	0.959	0.035
Step 2	0.018	NA	NA (NA / NA)	0.495	0.006	0.969	NA
Step 3a	0.015	4.361	0.207 (0.186 / 0.021)	0.486	0.006	0.961	0.031
Step 3b	0.009	1.176	0.062 (0.048 / 0.014)	0.315	0.005	0.955	0.816

Note: Oracle 1 knows both $J(\alpha_\gamma)$ and τ_0 and Oracle 2 knows only $J(\alpha_\gamma)$. Expectation (E) is calculated by the average of 1,000 iterations in each design. Note that $J(\alpha_\gamma) = 1$. 'NA' denotes 'Not Available' as the parameter is not estimated in the step. The estimation results for τ at the rows of Step 3a and Step 3b are based on the re-estimation of τ given estimates from Step 3a ($\widehat{\alpha}$) and Step 3b ($\widetilde{\alpha}$).

Table 2: Baseline Model: $\gamma = 0.5$

	Excess Risk	E$[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c}$)	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop.
$n = 200$							
Oracle 1	0.008	NA	0.012 (NA / NA)	0.288	NA	NA	NA
Oracle 2	0.018	NA	0.012 (NA / NA)	0.465	0.011	0.948	NA
Step 1	0.040	5.731	0.279 (0.245 / 0.034)	0.723	0.011	0.950	0.015
Step 2	0.036	NA	NA (NA / NA)	0.729	0.011	0.946	NA
Step 3a	0.039	5.776	0.272 (0.239 / 0.033)	0.717	0.011	0.947	0.017
Step 3b	0.040	2.364	0.182 (0.155 / 0.027)	0.702	0.011	0.929	0.428
$n = 400$							
Oracle 1	0.004	NA	0.006 (NA / NA)	0.201	NA	NA	NA
Oracle 2	0.008	NA	0.006 (NA / NA)	0.337	0.005	0.956	NA
Step 1	0.022	6.055	0.144 (0.128 / 0.017)	0.512	0.005	0.953	0.020
Step 2	0.020	NA	NA (NA / NA)	0.509	0.005	0.950	NA
Step 3a	0.019	6.056	0.142 (0.126 / 0.017)	0.517	0.005	0.947	0.020
Step 3b	0.018	2.250	0.061 (0.054 / 0.007)	0.460	0.005	0.949	0.649

Note: $J(\alpha_\gamma) = 2$. See the note below Table 1 for other notation.

Table 3: Baseline Model: $\gamma = 0.75$

	Excess Risk	$\mathbb{E}[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c}$)	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop.
$n = 200$							
Oracle 1	0.008	NA	0.015 (NA / NA)	0.324	NA	NA	NA
Oracle 2	0.016	NA	0.015 (NA / NA)	0.508	0.011	0.941	NA
Step 1	0.043	6.056	0.352 (0.310 / 0.042)	0.769	0.012	0.930	0.024
Step 2	0.036	NA	NA (NA / NA)	0.787	0.013	0.911	NA
Step 3a	0.036	6.045	0.349 (0.308 / 0.042)	0.782	0.013	0.911	0.024
Step 3b	0.029	2.160	0.232 (0.188 / 0.044)	0.629	0.012	0.925	0.688
$n = 400$							
Oracle 1	0.004	NA	0.007 (NA / NA)	0.218	NA	NA	NA
Oracle 2	0.008	NA	0.007 (NA / NA)	0.354	0.005	0.952	NA
Step 1	0.018	6.007	0.169 (0.150 / 0.020)	0.538	0.005	0.962	0.013
Step 2	0.019	NA	NA (NA / NA)	0.571	0.005	0.944	NA
Step 3a	0.019	6.032	0.169 (0.149 / 0.020)	0.548	0.005	0.942	0.016
Step 3b	0.010	2.128	0.052 (0.041 / 0.011)	0.367	0.005	0.953	0.860

Note: $J(\alpha_\gamma) = 2$. See the note below Table 1 for other notation.

We report the excess risk, the average number of parameters selected, $\mathbb{E}[J(\widehat{\alpha})]$, and the sum of the mean squared error of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}$ / $\widehat{\alpha}_{J_0^c}$). For each sample, the excess risk is calculated by the simulation, $S^{-1} \sum_{s=1}^{S} \left[\rho(Y_s, X_s^T(\widehat{\tau})\widehat{\alpha}) - \rho(Y_s, X_s^T(\tau_0)\alpha_0) \right]$, where $S = 10{,}000$ is the number of simulations; then we report the average value of 1,000 replications. Similarly, we also calculate prediction errors by the simulation, $\left(S^{-1} \sum_{s=1}^{S} \left(X_s^T(\widehat{\tau})\widehat{\alpha} - X_s^T(\tau_0)\alpha_\gamma \right)^2 \right)^{1/2}$, and report the average value.

We also report the root-mean-squared error (RMSE) and the coverage probability of the 95% confidence interval of $\widehat{\tau}$ (C. Prob. of $\widehat{\tau}$). The confidence intervals for τ_0 are calculated by simulating the two-sided compound Poisson process in Theorem 4.3 by adopting the approach proposed by Li and Ling (2012). The details are provided in Section A. Li and Ling (2012) showed that it is valid to simulate the compound poisson process by simulating the poisson process and the compounding factors from empirical distributions separately in the context of least squares estimation. We build on their suggestion and modify their procedure to quantile regression. We did not prove a formal justification for our procedure in this paper; however, it seems working well in simulations. It is an interesting topic for future research.

Note that the root-mean-squared error of $\widehat{\tau}$ and the coverage probability of the confidence

interval at the rows of Step 3a and Step 3b in the tables are estimation results of updated $\widehat{\tau}$: we re-estimate τ as in Step 2 using $(\widehat{U}_i, \widehat{\alpha})$ and $(\widetilde{U}_i, \widetilde{\alpha})$ from Step 3a and Step 3b instead of $(\breve{U}_i, \breve{\alpha})$. Finally, we also report the oracle proportion (Oracle Prop.), namely the ratio of the correct model selection out of 1,000 replications.

Overall, the simulation results confirm the asymptotic theory developed in the previous sections. First, these results show the advantage of quantile regression models over the existing mean regression models with a change point, e.g. Lee et al. (2016). The proposed estimator (Step 3b) selects different nonzero coefficients at different quantile levels. The estimator in Lee et al. (2016) cannot detect these heterogeneous models. In general the proposed estimators show better performance for heteroskedastic designs and for the fat-tail error distributions as will be discussed in detail below. Second, when we look at the finite sample performance of the proposed estimators in Step 3, their prediction errors are within a reasonable bound from those of Oracles 1 and 2. Recall that we estimate models with 250 times or 500 times more regressors in each design. Third, the root-mean-squared error of $\widehat{\tau}$ decreases quickly and confirms the super-consistency result of $\widehat{\tau}$. Fourth, the coverage probabilities of the confidence interval are close to 95%, especially when $n = 400$. Thus, we recommend practitioners to use $\widehat{\tau}$ in Step 2 or the re-estimated version of it based on the estimates from Step 3a or Step 3b. Finally, the oracle proportion of Step 3b is quite satisfactory and confirms our results in model selection consistency.

6.1 Comparison with Mean Regression with a Change Point

Table 4 compares the performance of the proposed estimator with that of the mean regression method in Lee et al. (2016). For the purpose of direct comparison between mean and median regression models, the tuning parameter λ is fixed to be the same as that in Step 1 from median regression. We consider three different simulation designs at $\gamma = 0.5$ with $n = 200$. The first model is a homoskedastic model by setting $\xi_{01} = (1, 0, \ldots, 0)$ in the baseline design. The second model is the same as the heteroskedastic median regression in

Table 2. The third model is a fat-tail model, where U_i follows a Cauchy distribution with a scale parameter 0.25 while keeping the heteroskedastic design as the second model. The mean regression method shows slight over-selection but its performance looks reasonable in the homoskedastic model. However, the method in Lee et al. (2016) is not robust to the heteroskedastic errors, which we can observe in Panel B of Table 4. Furthermore, it cannot detect different nonzero coefficients at different quantile levels while the quantile method shows such a result in Table 1. Finally, the quantile method works well when the error distribution follows a Cauchy distribution in Panel C of Table 4. However, the mean regression method performs poorly with a Cauchy error distribution as the conditional mean function is not well-defined in this case.

Table 4: Comparison between mean and median regression models with a change point

Panel A—Homoskedastic Model: $\gamma = 0.5$ and $n = 200$

	$\mathrm{E}[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ $(\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c})$	Pred. Er.	RMSE of $\widehat{\tau}$	Oracle Prop.
Oracle 1	NA	0.000 (NA / NA)	0.056	NA	NA
Oracle 2	NA	0.000 (NA / NA)	0.199	0.003	NA
Step 1	5.919	0.011 (0.010 / 0.001)	0.259	0.003	0.026
Step 2	NA	NA (NA / NA)	0.248	0.003	NA
Step 3a	5.900	0.011 (0.010 / 0.001)	0.257	0.003	0.024
Step 3b	2.001	0.001 (0.001 / 0.000)	0.213	0.003	0.999
Mean Reg	8.162	0.010 (0.008 / 0.001)	0.256	0.003	0.000

Panel B—Heteroskedastic Model: $\gamma = 0.5$ and $n = 200$

	$\mathrm{E}[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ $(\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c})$	Pred. Er.	RMSE of $\widehat{\tau}$	Oracle Prop.
Oracle 1	NA	0.012 (NA / NA)	0.288	NA	NA
Oracle 2	NA	0.012 (NA / NA)	0.465	0.011	NA
Step 1	5.731	0.279 (0.245 / 0.034)	0.723	0.011	0.015
Step 2	NA	NA (NA / NA)	0.729	0.011	NA
Step 3a	5.776	0.272 (0.239 / 0.033)	0.717	0.011	0.017
Step 3b	2.364	0.182 (0.155 / 0.027)	0.702	0.011	0.428
Mean Reg	93.550	2.537 (0.326 / 2.211)	1.572	0.011	0.000

Panel C—Fat-tail Model: $\gamma = 0.5$, $U_i \sim Cauchy(0.25)$ and $n = 200$

	$\mathrm{E}[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ $(\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c})$	Pred. Er.	RMSE of $\widehat{\tau}$	Oracle Prop.
Oracle 1	NA	0.005(NA / NA)	0.185	NA	NA
Oracle 2	NA	0.005(NA / NA)	0.392	0.011	NA
Step 1	5.843	0.148 (0.131 / 0.017)	0.566	0.011	0.022
Step 2	NA	NA (NA / NA)	0.576	0.011	NA
Step 3a	5.806	0.143 (0.126 / 0.017)	0.575	0.011	0.019
Step 3b	2.582	0.074 (0.066 / 0.008)	0.575	0.011	0.483
Mean Reg	218.991	5.55×10^6 (5.45×10^3 / 5.50×10^6)	137.985	0.221	0.000

6.2 When There Is No Change Point

Table 5 shows the performance of the estimator when there does not exist any change point. We use the baseline design with $\gamma = 0.75$ and set $\delta = (0, \ldots, 0)$. As we are interested in the performance of $\widehat{\delta}$, we report the average number of parameters selected in $\widehat{\delta}$, the MSE of $\widehat{\delta}$, and the proportion of detecting no-change point (No-change Prop.). As predicted by the theory, all measures on $\widehat{\delta}$ indicate that the estimator (Step 3b) detects no-change point models quite well. Both $\mathbb{E}[J(\widehat{\delta})]$ and MSE of $\widehat{\delta}$ are quite low and no-change proportion is high. We can also observe much improvement in these measure when the sample size increases from $n = 200$ to $n = 400$.

Table 5: No Change Point: $\gamma = 0.75$, $\delta_\gamma = 0$

	Excess Risk	$\mathbb{E}[J(\widehat{\alpha})]$	$\mathbb{E}[J(\widehat{\delta})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\delta}$	Pred. Er.	No-change Prop.	Oracle Prop.
$n = 200$								
Oracle 1	0.004	NA	NA	0.002	NA	0.196	NA	NA
Oracle 2	0.004	NA	NA	0.002	NA	0.196	NA	NA
Step 1	0.030	4.796	1.149	0.228	0.006	0.618	0.221	0.008
Step 2	0.024	NA	NA	NA	NA	0.617	NA	NA
Step 3a	0.026	4.915	1.309	0.226	0.008	0.602	0.142	0.008
Step 3b	0.017	1.520	0.334	0.178	0.007	0.436	0.722	0.541
$n = 400$								
Oracle 1	0.002	NA	NA	0.001	NA	0.143	NA	NA
Oracle 2	0.002	NA	NA	0.001	NA	0.143	NA	NA
Step 1	0.015	4.933	1.137	0.126	0.003	0.451	0.223	0.013
Step 2	0.014	NA	NA	NA	NA	0.449	NA	NA
Step 3a	0.015	5.042	1.301	0.124	0.004	0.440	0.123	0.010
Step 3b	0.005	1.208	0.141	0.037	0.002	0.197	0.867	0.805

Note: $J(\alpha_\gamma) = 1$ and $J(\delta_\gamma) = 0$.

6.3 When the Minimal Signal in δ is Low

In this subsection, we consider the case when the model contains low *minimal* signals in δ. Specifically, we consider the median regression model and set $\beta_{0.5} = (0, 0.34, 0, \ldots, 0)$ and $\delta_{0.5} = (0, 1, 1/2, 1/4, 1/8, 1/16, 0 \ldots, 0)$. Table 6 reports simulation results in this design. Note that the simulation design in Table 6 is the same as that reported in Table 2 except that $\delta_{0.5} = (0, 1, 0 \ldots, 0)$ in Table 2. Therefore, we may view that the simulation design in Table 2 satisfies the minimum signal condition, whereas that of this subsection does not.

The simulation results in Table 6 are consistent with asymptotic theory in Section 4 and remarks in Section 2.2 comparing estimators in step 3. The step 3b estimator performs better than the step 3a estimator in Table 2, but it performs worse in Table 6. Also note that the oracle proportion is zero for the step 3b estimator, which is expected given low signals in coefficients. Finally, it is important to note that the performance of the estimators of τ_0 is good in terms of the MSE in the presence of low signals in δ. The coverage probability of the confidence interval is much higher than the nominal level, which was not observed in previous simulations. Since the MSE and coverage probability between the infeasible oracle 2 estimator and other estimators are very similar, we interpret that the over-coverage result is not driven by high-dimensionality of regressors and variable selection. Perhaps this is due to a larger number of coefficients to estimate for the oracle 2 estimator, compared to Table 2.

Table 6: When the minimal signal in δ is low

	Excess Risk	$\mathbb{E}[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}$ / $\widehat{\alpha}_{J_0^c}$)	Pred. Er.	MSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
$n = 200$							
Oracle 1	0.024	NA	0.729 (NA / NA)	0.522	NA	NA	NA
Oracle 2	0.037	NA	0.729 (NA / NA)	0.816	0.004	0.995	NA
Step 1	0.054	9.949	0.517 (0.414 / 0.104)	0.946	0.004	0.995	0.000
Step 2	0.054	NA	NA (NA / NA)	0.949	0.004	0.992	NA
Step 3a	0.056	9.923	0.517 (0.414 / 0.104)	0.903	0.004	0.991	0.000
Step 3b	0.062	3.327	1.293 (1.174 / 0.119)	1.002	0.004	0.990	0.000
$n = 400$							
Oracle 1	0.012	NA	0.339 (NA / NA)	0.365	NA	NA	NA
Oracle 2	0.017	NA	0.339 (NA / NA)	0.522	0.002	0.999	NA
Step 1	0.029	11.058	0.333 (0.275 / 0.058)	0.647	0.002	1.000	0.000
Step 2	0.029	NA	NA (NA / NA)	0.694	0.002	0.997	NA
Step 3a	0.027	11.067	0.332 (0.274 / 0.058)	0.678	0.002	0.998	0.000
Step 3b	0.032	3.574	0.648 (0.585 / 0.063)	0.691	0.002	0.999	0.000

6.4 Additional Simulation Results

We have carried out additional Monte Carlo experiments. For the sake of brevity, we only report main findings here and show full results in the appendices. In Appendix F, we report simulation results when the change point τ_0 and the distribution of Q_i vary. In particular, we consider three different distributions of Q_i: Uniform$[0, 1]$, $N(0, 1)$, and $\chi^2(1)$.

The change point parameter τ_0 varies over $0.3, 0.4, \ldots, 0.7$ quantiles of each Q_i distribution. We find that the performance of $\widehat{\tau}$ measured by the root-mean-squared error depends on the density of Q_i distribution, as is expected from asymptotic theory. For instance, it is quite uniform over different τ_0 when Q_i follows Uniform$[0, 1]$. However, when Q_i follows $N(0, 1)$ or $\chi^2(1)$, it performs better when τ_0 is located at a point with a higher density of Q_i distribution. Sensitivity analyses provided in Appendix G show that the main simulation results are robust when we make changes over the range between -15% and $+15\%$ of the suggested tuning parameter values.

In summary, the proposed estimation procedure works well in finite samples and confirms the theoretical results developed earlier. The simulation studies show some advantages of the proposed estimator over the existing mean regression method, e.g. Lee et al. (2016). It also detects no-change-point models well without any pre-test. The main qualitative results are not sensitive to different simulation designs on τ_0 and Q_i as well as to some variation on tuning parameter values.

7   Estimating a Change Point in Racial Segregation

As an empirical illustration, we investigate the existence of tipping in the dynamics of racial segregation using the dataset constructed by Card et al. (2008). They show that the neighborhood's white population decreases substantially when the minority share in the area exceeds a tipping point (or threshold point), using U.S. Census tract-level data. Lee et al. (2011) develop a test for the existence of threshold effects and apply their test to this dataset. Different from these existing studies, we consider a high-dimensional setup by allowing both possibly highly nonlinear effects of the main covariate (minority share in the neighborhood) and possibly higher-order interactions between additional covariates.

We build on the specifications used in Card et al. (2008) and Lee et al. (2011) to choose

the following median regression with a constant shift due to the tipping effect:

$$Y_i = g_0(Q_i) + \delta_0 1\{Q_i > \tau_0\} + X_i'\beta_0 + U_i, \tag{7.1}$$

where for census tract i, the dependent variable Y_i is the ten-year change in the neighborhood's white population, Q_i is the base-year minority share in the neighborhood, and X_i is a vector of six tract-level control variables and their various interactions depending on the model specification. Both Y_i and Q_i are in percentage terms. The basic six variables in X_i include the unemployment rate, the log of mean family income, the fractions of single-unit, vacant, and renter-occupied housing units, and the fraction of workers who use public transport to travel to work. The function $g(\cdot)$ is approximated by the cubic b-splines with 15 knots over equi-quantile locations, so the degrees of freedom are 19 including an intercept term. In our empirical illustration, we use the census-tract-level sample of Chicago whose base year is 1980.

In the first set of models, we consider possible interactions among the six tract-level control variables up to six-way interactions. Specifically, the vector X in the six-way interactions will be composed of the following 63 regressors,

$$\{X^{(1)}, \ldots, X^{(6)}, X^{(1)}X^{(2)}, \ldots, X^{(5)}X^{(6)}, \ldots, X^{(1)}X^{(2)}X^{(3)}X^{(4)}X^{(5)}X^{(6)}\},$$

where $X^{(j)}$ is the j-th element among those tract-level control variables. Note that the lower order interaction vector (e.g. two-way or three-way) is nested by the higher order interaction vector (e.g. three-way or four-way). The total number of regressors varies from 26 (19 from b-splines, 6 from X_i and $1\{Q_i > \tau\}$) when there is no interaction to 83 when there are full six-way interactions. In the next set of models, we add the square of each tract-level control variable and generate similar interactions up to six. In this case the total number of regressors varies from 32 to 2,529. For example, the number of regressors in the largest model consists of

#(b-spline basis) + #(indicator function) + #(interactions up to six-way out of 12) = 19 +

$1 + \sum_{k=1}^{6} \binom{12}{k} = 2,529$. This number is much larger than the sample size ($n = 1,813$).

Table 7 summarizes the estimation results at the 0.25, 0.5, and 0.75 quantiles, respectively. We report the total number of regressors in each model and the number of selected regressors in Step 3b. The change point τ is estimated by the grid search over 591 equi-spaced points in $[1, 60]$. The lower bound value 1% corresponds to the 1.6 sample percentile of Q_i and the upper bound value 60%, which is about the upper sample quartile of Q_i, is the same as one used in Card et al. (2008). In this empirical example, we report the estimates of τ_0 and the confidence intervals updated after Step 3b (that is, τ is re-estimated using the estimates of α_0 in Step 3b). If this estimate is different from the previous one in Step 2, then we repeat Step 3b and Step 2 until it converges.

The estimation results suggest several interesting points. First, at each quantile, the proposed method selects sparse representations in all model specifications even when the number of regressors is relatively large. Furthermore, the number of selected regressors does not grow rapidly when we increase the number of possible covariates. It seems that the set of selected covariates overlaps across different dictionaries at each quantile. See Appendix H for details on selected regressors. Second, the estimation results are different across different quantiles, indicating that there may exist heterogeneity in this application. The confidence intervals for τ_0 at the 0.25 quantile are quite tight in all cases and they provide convincing evidence of the tipping effect. If we look at the case of six-way interactions with 12 control variables, the estimated tipping point is 5.65% and the estimated jump size is -5.50%. However, this strong tipping effect becomes weaker at the 0.50 and 0.75 quantiles as shown either by wider confidence intervals or by the zero jump size, i.e. $\widehat{\delta} = 0$.

We now compare the estimation results from quantile regression with those from mean regression, which are reported in Table 8 (full estimation results are in Appendix H). We show two kinds of mean regression estimates: one with the untrimmed original data and the other with the trimmed data for which we drop top and bottom 5% observations based on $\{Y_i\}$. The estimated tipping points are the same between the two datasets but the estimated

Table 7: Estimation Results from Quantile Regression

	No. of Reg.	No. of Selected Reg. in Step 3b	$\widehat{\tau}$	CI for τ_0	$\widehat{\delta}$
$\gamma = 0.25$					
6 control variables					
No Interaction	26	17	5.65	[4.75, 6.17]	-4.07
Two-way Interaction	41	20	2.35	[1.00, 4.44]	-1.82
Three-way Interaction	61	24	2.35	[1.00, 4.15]	-2.19
Four-way Interaction	76	21	5.65	[4.69, 6.08]	-5.50
Five-way Interaction	82	22	2.45	[1.00, 4.93]	-1.55
Six-way Interaction	83	22	2.45	[1.00, 4.75]	-1.55
12 control variables					
No Interaction	32	17	5.65	[4.75, 6.17]	-4.07
Two-way Interaction	98	18	5.25	[3.55, 6.09]	-3.40
Three-way Interaction	318	22	5.25	[3.63, 5.94]	-3.61
Four-way Interaction	813	26	5.25	[3.79, 5.97]	-3.53
Five-way Interaction	1605	27	5.25	[4.57, 5.65]	-5.37
Six-way Interaction	2529	28	5.65	[4.96, 6.06]	-5.50
$\gamma = 0.50$					
6 control variables					
No Interaction	26	15	5.65	[1.67, 11.85]	-2.24
Two-way Interaction	41	17	5.05	[2.25, 7.46]	-2.63
Three-way Interaction	61	20	5.25	[4.22, 6.38]	-4.15
Four-way Interaction	76	19	5.05	[3.60, 7.00]	-3.14
Five-way Interaction	82	20	5.05	[1.23, 9.16]	-1.90
Six-way Interaction	83	20	5.05	[1.33, 9.39]	-1.90
12 control variables					
No Interaction	32	16	1.95	[0.77, 4.61]	-3.69
Two-way Interaction	98	21	6.75	[1.00, 45.57]	0.48
Three-way Interaction	318	25	4.05	[1.00, 13.15]	-0.97
Four-way Interaction	813	27	3.65	[1.00, 15.91]	-0.56
Five-way Interaction	1605	29	3.25	[1.00, 13.16]	-0.68
Six-way Interaction	2529	28	3.25	[1.00, 11.67]	-0.74
$\gamma = 0.75$					
6 control variables					
No Interaction	26	15	10.05	[9.37, 11.29]	-10.62
Two-way Interaction	41	14	NA	NA	0.00
Three-way Interaction	61	21	NA	NA	0.00
Four-way Interaction	76	18	NA	NA	0.00
Five-way Interaction	82	18	NA	NA	0.00
Six-way Interaction	83	18	NA	NA	0.00
12 control variables					
No Interaction	32	14	10.05	[8.44, 11.94]	-7.14
Two-way Interaction	98	20	NA	NA	0.00
Three-way Interaction	318	21	NA	NA	0.00
Four-way Interaction	813	25	NA	NA	0.00
Five-way Interaction	1605	28	NA	NA	0.00
Six-way Interaction	2529	24	NA	NA	0.00

Note: The sample size is $n = 1,813$. The parameter τ_0 is estimated by the grid search on the 591 equi-spaced points over $[1, 60]$. Both $\widehat{\tau}$ and the 95% confidence interval are based on re-estimation after Step 3b: that is, τ is estimated again using $(\widetilde{U}_i, \widetilde{\alpha})$ from Step 3b.

Table 8: Estimation Results from Mean Regression

	No. of Reg.	No. of Selected Reg.	$\widehat{\tau}$	$\widehat{\delta}$
6 Control Variables, Six-way Interaction				
Untrimmed	83	50	3.25	-16.14
Trimmed	83	41	3.25	-6.53
12 Control Variables, Six-way Interaction				
Untrimmed	2529	142	3.25	-15.55
Trimmed	2529	107	3.25	-5.19

Note: The sample size of the untrimmed original data is $n = 1,813$. The trimmed data drop top and bottom 5% observations based on $\{Y_i\}$ and the sample sizes decreases to $n = 1,626$. The parameter τ_0 is estimated by the grid search on the 591 equi-spaced points over $[1, 60]$. As in the simulation studies, the tuning parameters are set from Step 1 in median regression.

jump size is much larger with the original data. Figure 1 shows the fitted values over Q_i at the sample mean of the six basic covariates. They are from the model of six-way interactions with 12 control variables and the vertical line indicates the location of a tipping point. The left panel of Figure 1 compares the results between the mean and median regression results (without trimming the data) and the right panel shows the the interquartile range of the conditional distribution of Y_i as a function of Q_i given other regressors. It can be seen that the mean regression estimates are much more volatile around the tipping point than the median regression estimates, although the estimated tipping point is the same. In Figure 2, we compare the mean regression estimates with and without trimming. Removing observations with top and bottom 5% Y_i's stablize the estimates, thus demonstrating that the median regression estimates have the built-in feature that they are more stable with outliers of Y_i than the mean estimates. Finally, looking at the right panel of Figure 1, we can see that the 25 percentile of the conditional distribution drops at the tipping point of 5.65% but no such change at the 75% quantile. This shows that the quantile regression estimates can provide insights into *distributional* threshold effects in racial segregation.

In summary, this empirical example shows that the proposed method works well in the real empirical setup and is robust to outliers compared to the mean regression approach. The estimation results also confirm that there exists a tipping point in the racial segregation

Figure 1: Estimation Results: 12 Control Variables and Six-way Interaction
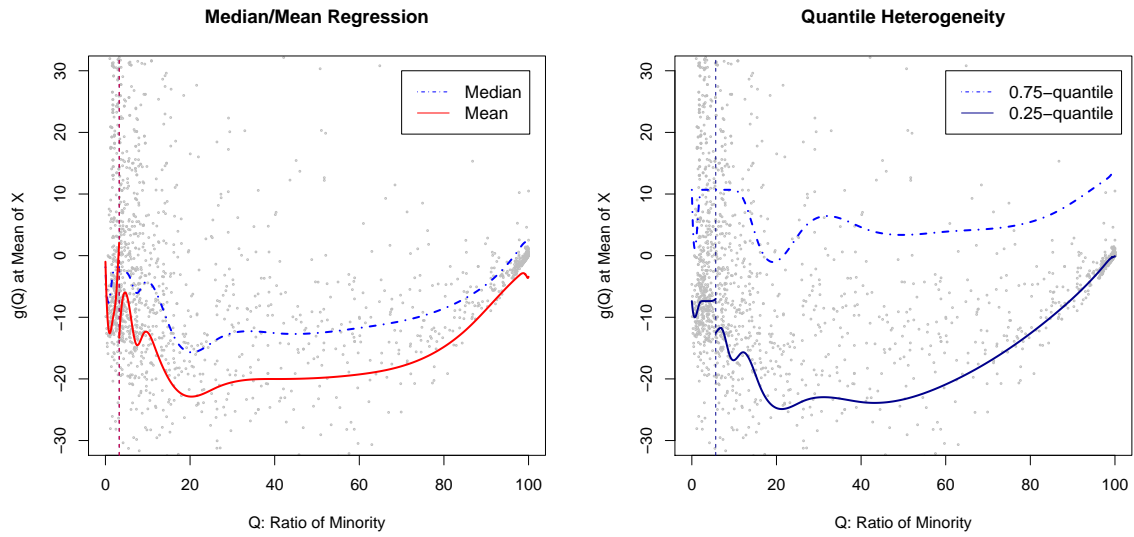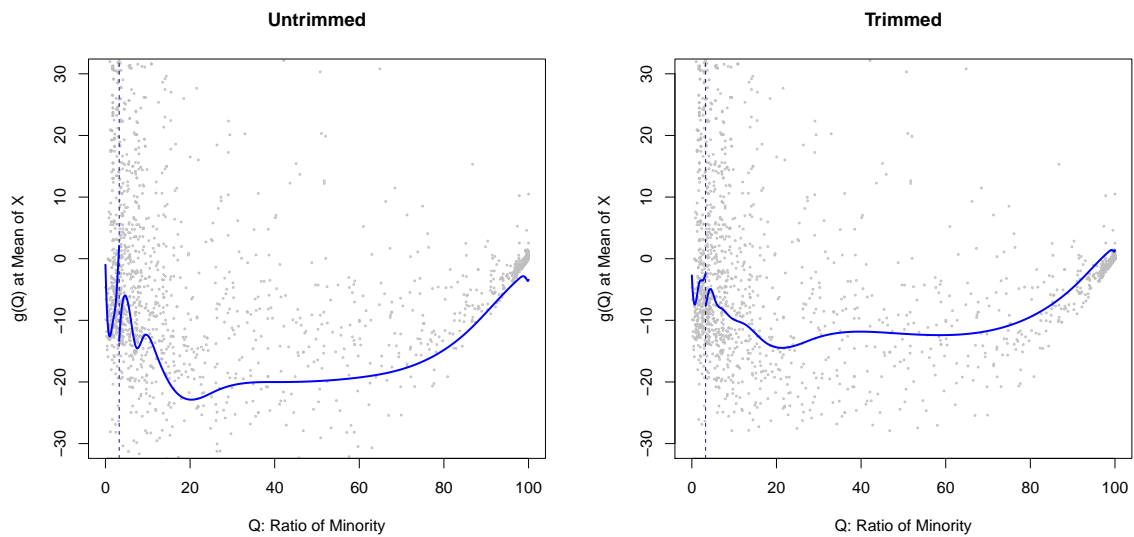

Median/Mean Regression

Quantile Heterogeneity

Figure 2: Estimation Results: Mean Regression with Untrimmed/Trimmed Data

Untrimmed

Trimmed

at the 0.25 quantile and that the tipping effect is heterogeneous over different quantiles.

8 Conclusions

In this paper, we have developed ℓ_1-penalized estimators of a high-dimensional quantile regression model with an unknown change point due to a covariate threshold. We have shown among other things that our estimator of the change point achieves an oracle property without relying on a perfect covariate selection, thereby avoiding the need for the minimum level condition on the signals of active covariates. We have illustrated the usefulness of our estimation methods via Monte Carlo experiments and an application to tipping in the racial segregation.

In a recent working paper, Leonardi and Bühlmann (2016) consider a high-dimensional *mean* regression model with multiple change points whose number may grow as the sample size increases. They have proposed a binary search algorithm to choose the number of change points. It is an important future research topic to develop a computationally efficient algorithm to detect multiple changes for high-dimensional quantile regression models.

Appendices

In Appendix A, we provide the algorithm of constructing the confidence interval for τ_0. In Appendix B, we provide a set of regularity assumptions to derive asymptotic properties of the proposed estimators in Sections 4 and 5. In Appendix C, we provide sufficient conditions for the identification of (α_0, τ_0) in (2.1) and show that an improved rate of convergence is possible for the excess risk by taking the second and third steps of estimation. To prove the theoretical results in the main text, we consider a general M-estimation framework that includes quantile regression as a special case. We provide high-level regularity conditions on the loss function in Appendix D. Under these conditions, we derive asymptotic properties and then we verify all the high level assumptions for the quantile regression model in Appendix E. Hence, our general results are of independent interest and can be applicable to other models, for example logistic regression models. Appendices F and G provide additional simulation results, and Appendix H gives additional results for the empirical example.

A The Algorithm of Constructing the Confidence Interval for τ_0

The detailed algorithm for constructing the confidence interval based on the Step 2 estimator is as follows:

1. Simulate two independent Poisson processes $N_1(-h)$ for $h < 0$ and $N_2(h)$ for $f > 0$ with the same jump rate $\widehat{f}_Q(\widehat{\tau})$ over $h \in [-\overline{H}n, \overline{H}n]$, where $f_Q(\cdot)$ is the pdf of Q, n is the sample size, and $\overline{H} > 0$ is a large constant. For estimating $f_Q(\cdot)$, we use the kernel density estimator with a normal density kernel and the rule-of-thumb bandwidth, $1.06 \cdot \min\{s, (Q_{0.75} - Q_{0.25})/1.34\} \cdot n^{-1/5}$, where s is the standard deviation of Q and $Q_{0.75} - Q_{0.25}$ is the interquartile range of Q. A Poisson process $N(h)$ is generated by the following algorithm:

(a) Set $h = 0$ and $k = 0$.

(b) Generate ϵ from the uniform distribution on $[0, 1]$.

(c) $h = h + [-(1/\widehat{f}_Q(\widehat{\tau})) \log(\epsilon)]$.

(d) If $h > n\overline{H}$, then stop and goto Step (f). Otherwise, set $k = k + 1$ and $h_k = h$.

(e) Repeat Steps (b)–(d).

(f) The algorithm generates $\{h_k\}$ for $k = 1, \ldots, \overline{K}$. Transform it into the Poisson process $N(h) \equiv \sum_{k=1}^{\overline{K}} 1\{h_k \leq h\}$ for $h \in [0, n\overline{H}]$.

2. Using the residuals $\{\breve{U}_i\}$ and the estimate $\breve{\delta}$ from Step 1, simulate ρ_{1j} for $j = 1, \ldots, N_1(-h)$ from the empirical distribution of $\{\dot{\rho}(\breve{U}_i - X_i^T \breve{\delta}) - \dot{\rho}(\breve{U}_i)\}_{i \leq n}$; simulate ρ_{2j} for $j = 1, \ldots, N_2(h)$ from the empirical distribution of $\{\dot{\rho}(\breve{U}_i + X_i^T \breve{\delta}) - \dot{\rho}(\breve{U}_i)\}_{i \leq n}$. Here $\dot{\rho}(t) \equiv t(\gamma - 1\{t \leq 0\})$ is the check function as defined in Section 4.

3. Recall that
$$M(h) \equiv \sum_{i=1}^{N_1(-h)} \rho_{1i} 1\{h < 0\} + \sum_{i=1}^{N_2(h)} \rho_{2i} 1\{h \geq 0\}$$
from Section 4. Construct the function $M(\cdot)$ for $h \in [-\overline{H}n, \overline{H}n]$ using values from Steps 1–3 above. Find the smallest minimizer h of $M(\cdot)$.

4. Repeat Steps 1–4 above and generate $\{h_1, \ldots, h_B\}$.

5. Construct the 95% confidence interval of $\widehat{\tau}$ from the empirical distribution of $\{h_b\}$ by $[\widehat{\tau} + h_{0.025}/n, \widehat{\tau} + h_{0.975}/n]$, where $h_{0.025}$ and $h_{0.975}$ are 2.5 and 97.5 percentiles of $\{h_b\}$, respectively.

It is straightforward to modify the algorithm above for the confidence intervals with Step 3a and Step 3b estimators. We set $\overline{H} = 0.5$, and $B = 1,000$ in this simulation studies.

B Assumptions for Oracle Properties

In this section, we list a set of assumptions that will be useful to derive asymptotic properties of the proposed estimators in Sections 4 and 5. In the following, we divide our

discussions into two important cases: (i) $\delta_0 \neq 0$ and τ_0 is identified, and (ii) $\delta_0 = 0$ and thus τ_0 is not identified. The asymptotic properties are derived under both cases. Note that such a distinction is only needed for presenting our theoretical results. In practice, we do not need to know whether $\delta_0 = 0$ or not.

Assumption 2 (Underlying Distribution). *(i) The conditional distribution $Y|X,Q$ has a continuously differentiable density function $f_{Y|X,Q}(y|x,q)$ with respect to y, whose derivative is denoted by $\tilde{f}_{Y|X,Q}(y|x,q)$.*

(ii) There are constants $C_1, C_2 > 0$ such that for all (y,x,q) in the support of (Y,X,Q),

$$|\tilde{f}_{Y|X,Q}(y|x,q)| \leq C_1, \quad f_{Y|X,Q}(x(\tau_0)^T \alpha_0 |x,q) \geq C_2.$$

(iii) When $\delta_0 \neq 0$, $\Gamma(\tau, \alpha_0)$ is positive definite uniformly in a neighborhood of τ_0, where

$$\Gamma(\tau, \alpha_0) \equiv \frac{\partial^2 \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_{0J})]}{\partial \alpha_J \partial \alpha_J^T} = \mathbb{E}[X_J(\tau) X_J(\tau)^T f_{Y|X,Q}(X(\tau)^T \alpha_0 |X,Q)].$$

When $\delta_0 = 0$, the matrix $\mathbb{E}[X_{J(\beta_0)} X_{J(\beta_0)}^T f_{Y|X,Q}(X_{J(\beta_0)}^T \beta_{0J(\beta_0)} |X,Q)]$ is positive definite.

Conditions (i) and (ii) are standard assumptions for quantile regression models. To follow the notation in condition (iii), recall that α_J denotes the subvector of α whose indices are in $J(\alpha_0)$. Expressions $X_J(\tau)$, $X_{J(\beta_0)}$, α_{0J} and $\beta_{0J(\beta_0)}$ can be understood similarly. Condition (iii) is a weak condition that imposes non-singularity of the Hessian matrix of the population objective function uniformly in a neighborhood of τ_0 in case of $\delta_0 \neq 0$. This condition reduces to the usual non-singularity condition when $\delta_0 = 0$.

B.1   Compatibility Conditions

We now make an assumption that is an extension of the well-known *compatibility condition* (see Bühlmann and van de Geer (2011), Chapter 6). In particular, the following

condition is a uniform-in-τ version of the compatibility condition. Recall that for a $2p$ dimensional vector α, we use α_J and α_{J^c} to denote its subvectors formed by indices in $J(\alpha_0)$ and $\{1, ..., 2p\} \setminus J(\alpha_0)$, respectively.

Assumption 3 (Compatibility Condition). *(i) When $\delta_0 \neq 0$, there is a neighborhood $\mathcal{T}_0 \subset \mathcal{T}$ of τ_0, and a constant $\phi > 0$ such that for all $\tau \in \mathcal{T}_0$ and all $\alpha \in \mathbb{R}^{2p}$ satisfying $|\alpha_{J^c}|_1 \leq 5|\alpha_J|_1$,*

$$\phi|\alpha_J|_1^2 \leq s\alpha^T \mathbb{E}[X(\tau)X(\tau)^T]\alpha. \tag{B.1}$$

(ii) When $\delta_0 = 0$, there is a constant $\phi > 0$ such that for all $\tau \in \mathcal{T}$ and all $\alpha \in \mathbb{R}^{2p}$ satisfying $|\alpha_{J^c}|_1 \leq 4|\alpha_J|_1$,

$$\phi|\alpha_J|_1^2 \leq s\alpha^T \mathbb{E}[X(\tau)X(\tau)^T]\alpha. \tag{B.2}$$

Assumption 3 requires that the compatibility condition hold uniformly in τ over a neighbourhood of τ_0 when $\delta_0 \neq 0$ and over the entire parameter space $\mathcal{T}$ when $\delta_0 = 0$. Note that this assumption is imposed on the population covariance matrix $\mathbb{E}[X(\tau)X(\tau)^T]$; thus, a simple sufficient condition of Assumption 3 is that the smallest eigenvalue of $\mathbb{E}[X(\tau)X(\tau)^T]$ is bounded away from zero uniformly in τ. Even if $p > n$, the population covariance can still be strictly positive definite while the sample covariance is not.

B.2 Restricted Nonlinearity Conditions

In this subsection, we make an assumption called a *restricted nonlinear condition* to deal with the quantile loss function. We extend condition D.4 in Belloni and Chernozhukov (2011) to accommodate the possible existence of the unknown threshold in our model (specifically, a uniform-in-τ version of the restricted nonlinear condition as in the compatibility condition).

Note that when $Q \leq \tau_0$, $X(\tau_0)^T\alpha_0 = X^T\beta_0$, while when $Q > \tau_0$, $X(\tau_0)^T\alpha_0 = X^T\theta_0$, where $\theta_0 \equiv \beta_0 + \delta_0$. Hence we define the "prediction balls" with radius r and corresponding

centers as follows:

$$\mathcal{B}(\beta_0, r) = \{\beta \in \mathcal{B} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}] \leq r^2\},$$

$$\mathcal{G}(\theta_0, r) = \{\theta \in \mathcal{G} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}] \leq r^2\},$$

(B.3)

where $\mathcal{B}$ and $\mathcal{G}$ are parameter spaces for β_0 and θ_0, respectively. To deal with the case that $\delta_0 = 0$, we also define

$$\tilde{\mathcal{B}}(\beta_0, r, \tau) = \{\beta \in \mathcal{B} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau\}] \leq r^2\},$$

$$\tilde{\mathcal{G}}(\beta_0, r, \tau) = \{\theta \in \mathcal{G} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}] \leq r^2\}.$$

(B.4)

Assumption 4 (Restricted Nonlinearity). *The following holds for the constants C_1 and C_2 defined in Assumption 2 (ii).*

(i) When $\delta_0 \neq 0$, there exists a constant $r_{QR}^* > 0$ such that

$$\inf_{\beta \in \mathcal{B}(\beta_0, r_{QR}^*), \beta \neq \beta_0} \frac{\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \leq \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \leq \tau_0\}]} \geq r_{QR}^* \frac{2C_1}{3C_2} > 0$$

(B.5)

and that

$$\inf_{\theta \in \mathcal{G}(\theta_0, r_{QR}^*), \theta \neq \theta_0} \frac{\mathbb{E}[|X^T(\theta - \theta_0)|^2 1\{Q > \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\theta - \theta_0)|^3 1\{Q > \tau_0\}]} \geq r_{QR}^* \frac{2C_1}{3C_2} > 0.$$

(B.6)

(ii) When $\delta_0 = 0$, there exists a constant $r_{QR}^* > 0$ such that

$$\inf_{\tau \in \mathcal{T}} \inf_{\beta \in \tilde{\mathcal{B}}(\beta_0, r_{QR}^*, \tau), \beta \neq \beta_0} \frac{\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \leq \tau\}]^{3/2}}{\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \leq \tau\}]} \geq r_{QR}^* \frac{2C_1}{3C_2} > 0$$

(B.7)

and that

$$\inf_{\tau \in \mathcal{T}} \inf_{\theta \in \tilde{\mathcal{G}}(\beta_0, r_{QR}^*, \tau), \beta \neq \beta_0} \frac{\mathbb{E}[|X^T(\theta - \theta_0)|^2 1\{Q > \tau\}]^{3/2}}{\mathbb{E}[|X^T(\theta - \theta_0)|^3 1\{Q > \tau\}]} \geq r_{QR}^* \frac{2C_1}{3C_2} > 0.$$

(B.8)

Remark B.1. As pointed out by Belloni and Chernozhukov (2011), if $X^T c$ follows a log-concave distribution conditional on Q for any nonzero c (e.g. if the distribution of X is multivariate normal), then Theorem 5.22 of Lovász and Vempala (2007) and the Hölder inequality imply that for all $\alpha \in \mathcal{A}$,

$$\mathbb{E}[|X(\tau_0)^T (\alpha - \alpha_0)|^3 | Q] \leq 6 \left\{ \mathbb{E}[\{X(\tau_0)^T (\alpha - \alpha_0)\}^2 | Q] \right\}^{3/2},$$

which provides a sufficient condition for Assumption 4. On the other hand, this assumption can hold more generally since equations (B.5)-(B.8) in Assumption 4 need to hold only locally around true parameters α_0.

B.3 Additional Assumptions When $\delta_0 \neq 0$

We first describe the additional conditions on the distribution of (X, Q).

Assumption 5 (Additional Conditions on the Distribution of (X, Q)). *Assume $\delta_0 \neq 0$. In addition, there exists a neighborhood $\mathcal{T}_0 \subset \mathcal{T}$ of τ_0 that satisfies the following.*

(i) Q has a density function $f_Q(\cdot)$ that is continuous and bounded away from zero on $\mathcal{T}_0$.

(ii) Let $\tilde{X}$ denote all the components of X excluding Q in case that Q is an element of X. The conditional distribution of Q given $\tilde{X}$ has a density function $f_{Q|\tilde{X}}(q|\tilde{x})$ that is bounded uniformly in both $q \in \mathcal{T}_0$ and $\tilde{x}$.

(iii) There exists $M_3 > 0$ such that $M_3^{-1} \leq \mathbb{E}[(X^T \delta_0)^2 | Q = \tau] \leq M_3$ for all $\tau \in \mathcal{T}_0$.

Condition (i) implies that $\mathbb{P}\{|Q - \tau_0| < \varepsilon\} > 0$ for any $\varepsilon > 0$, and condition (ii) requires that the conditional density of Q given $\tilde{X}$ be uniformly bounded. When τ_0 is identified, we require δ_0 to be considerably different from zero. This requirement is given in condition (iii). Note that this condition is concerned with $\mathbb{E}[(X^T \delta_0)^2 | Q = \tau]$, which is an important quantity to develop asymptotic results when $\delta_0 \neq 0$. Note that condition (iii) is a local

condition with respect to τ in the sense that it has to hold only locally in a neighborhood of τ_0.

The following additional moment conditions are useful to derive our theoretical results.

Assumption 6 (Moment Bounds). *(i) There exist finite positive constants $\widetilde{C}$ and r such that for all $\beta \in \mathcal{B}(\beta_0, r)$ and for any $\theta \in \mathcal{G}(\theta_0, r)$,*

$$\mathbb{E}[|X^T(\beta - \beta_0)|1\{Q > \tau_0\}] \leq \widetilde{C}\, \mathbb{E}[|X^T(\beta - \beta_0)|1\{Q \leq \tau_0\}],$$

$$\mathbb{E}[|X^T(\theta - \theta_0)|1\{Q \leq \tau_0\}] \leq \widetilde{C}\, \mathbb{E}[|X^T(\theta - \theta_0)|1\{Q > \tau_0\}].$$

(ii) There exist finite positive constants M, r and the neighborhood $\mathcal{T}_0$ of τ_0 such that

$$\mathbb{E}\left[(X^T[(\theta - \beta) - (\theta_0 - \beta_0)])^2 \big| Q = \tau\right] \leq M,$$

$$\mathbb{E}[|X^T(\beta - \beta_0)| \big| Q = \tau] \leq M,$$

$$\mathbb{E}[|X^T(\theta - \theta_0)| \big| Q = \tau] \leq M,$$

$$\sup_{\tau \in \mathcal{T}_0 : \tau > \tau_0} \mathbb{E}\left[|X^T(\beta - \beta_0)|\frac{1\{\tau_0 < Q \leq \tau\}}{(\tau - \tau_0)}\right] \leq M\mathbb{E}[|X^T(\beta - \beta_0)|1\{Q \leq \tau_0\}],$$

$$\sup_{\tau \in \mathcal{T}_0 : \tau < \tau_0} \mathbb{E}\left[|X^T(\theta - \theta_0)|\frac{1\{\tau < Q \leq \tau_0\}}{(\tau_0 - \tau)}\right] \leq M\mathbb{E}[|X^T(\theta - \theta_0)|1\{Q > \tau_0\}],$$

uniformly in $\beta \in \mathcal{B}(\beta_0, r)$, $\theta \in \mathcal{G}(\theta_0, r)$ and $\tau \in \mathcal{T}_0$.

Remark B.2. Condition (i) requires that Q have non-negligible support on both sides of τ_0. This condition can be viewed as a rank condition for identification of α_0. In the standard threshold model with a fixed dimension, our condition is trivially satisfied by the rank condition such that both $\mathbb{E}[XX^T1\{Q \leq \tau_0\}]$ and $\mathbb{E}[XX^T1\{Q > \tau_0\}]$ are positive definite (see e.g. Chan (1993) or Hansen (2000)). If the rank condition fails, the regression coefficient may not be identified and thus affecting the identification of the change point. In the high-dimensional setup, it is undesirable to impose the same rank condition due to the high-dimensionality. Instead, we replace it with condition (i). Condition (ii) requires the

boundedness and certain smoothness of the conditional expectation functions $\mathbb{E}[(X^T[(\theta - \beta) - (\theta_0 - \beta_0)])^2|Q = \tau]$, $\mathbb{E}[|X^T(\beta - \beta_0)||Q = \tau]$, and $\mathbb{E}[|X^T(\theta - \theta_0)||Q = \tau]$, and prohibits degeneracy in one regime. The last two inequalities in condition (ii) are satisfied if

$$\frac{\mathbb{E}\left[\left|X^T\beta\right||Q = \tau\right]}{\mathbb{E}\left[|X^T\beta|\right]} \le M$$

for all $\tau \in \mathcal{T}_0$ and for all β satisfying $0 < \mathbb{E}\left|X^T\beta\right| \le c$ for some small $c > 0$.

C   Additional Theoretical Results

In this part of the appendix, we consider the identification of (α_0, τ_0) in (2.1) and show that an improved rate of convergence is possible for the excess risk by taking the second and third steps of estimation.

C.1   Identification

The following theorem establishes the identification of (α_0, τ_0) in (2.1).

Theorem C.1 (Identification).   *(i) Assume that $\delta_0 \ne 0$ and that the γ-th conditional quantile of Y given X and Q is uniquely given as*

$$\text{Quantile}_{Y|X,Q}(\tau|X = x, Q = q) = x^T\beta_0 + x^T\delta_0 1\{q > \tau_0\}. \tag{C.1}$$

(ii) The distribution of Q is absolutely continuous with respect to Lebesgue measure.

(iii) $\tau_0 \in \mathcal{T} \equiv [\underline{\tau}, \overline{\tau}]$, which is contained in a strict interior of the support of Q.

(iv) For any $\tau_1 \in \mathcal{T}$ satisfying $\tau_1 < \tau_0$, we have that $P(\tau_1 < Q \le \tau_0) > 0$; for any $\tau_2 \in \mathcal{T}$ satisfying $\tau_2 > \tau_0$, $P(\tau_0 < Q \le \tau_2) > 0$.

(v) For every $\tau \in \mathcal{T}$, we have that $\inf_{q\in[\underline{\tau},\overline{\tau}]} \lambda_{\min}\{\mathbb{E}(X(\tau)X(\tau)^T|Q = q)\} > 0$, where $X(\tau) \equiv (X^T, X^T1\{Q > \tau\})^T$.

42

(vi) $\inf_{q \in [\underline{\tau}, \bar{\tau}]} \mathbb{E}((X^T \delta_0)^2 | Q = q) > 0.$

Then (α_0, τ_0) *is identified.*

Theorem C.1 establishes sufficient conditions under which α_0 and τ_0 are identified. Conditions (i)-(v) in Theorem C.1 are standard. The non-singularity condition (v) is uniform in $\tau \in \mathcal{T}$ and can be viewed as a natural extension of the usual rank condition in the linear model. Condition (vi) is a condition that imposes that the model is well separated from the case that there is no change point in the model.

Proof of Theorem C.1. Since the conditional quantile function is uniquely given as (C.1), it suffices to show that

$$X(\tau)^T \alpha = X(\tau_0)' \alpha_0 \text{ a.s.} \iff \alpha = \alpha_0 \text{ and } \tau = \tau_0.$$

To begin with, write, assuming $\tau \leq \tau_0$,

$$
\begin{aligned}
\mathcal{D}(\alpha, \tau) &\equiv X(\tau)^T \alpha - X(\tau_0)^T \alpha_0 \\
&= X^T (\beta - \beta_0) + X^T (\delta - \delta_0) \mathbf{1}\{Q > \tau\} + X^T \delta_0 \mathbf{1}\{\tau < Q \leq \tau_0\}.
\end{aligned}
\tag{C.2}
$$

Now suppose $\mathcal{D}(\alpha, \tau)$ in (C.2) is zero a.s. Then, it is also zero on the following event E:

$$E \equiv \{\mathbf{1}\{\tau < Q \leq \tau_0\} = 0\} = \{Q \notin (\tau, \tau_0]\}.
\tag{C.3}$$

on the other hand, $P(E) > 0$ because $P(E) = P(Q \notin (\tau, \tau_0]) \geq P(Q > \tau_0) > 0$. However, on event E,

$$\mathcal{D}(\alpha, \tau) = X^T (\beta - \beta_0) + X^T (\delta - \delta_0) \mathbf{1}\{Q > \tau\} = 0 \text{ a.s.}$$

Thus, we have that

$$X(\tau)^T (\alpha - \alpha_0) \mathbf{1}_E = 0 \text{ a.s.}$$

43

This is equivalent to

$$\mathbb{E}\{[X(\tau)^T(\alpha - \alpha_0)]^2 1_E\} = \mathbb{E}\{\mathbb{E}([X(\tau)^T(\alpha - \alpha_0)]^2|Q)1_E\} = 0.$$

However, we have that

$$
\begin{aligned}
0 &= \mathbb{E}\{\mathbb{E}([X(\tau)^T(\alpha - \alpha_0)]^2|Q)1_E\} \\
&\geq \inf_{q \in [\underline{\tau}, \bar{\tau}]} \mathbb{E}\{[X(\tau)^T(\alpha - \alpha_0)]^2|Q = q\}P(E) \\
&\geq \inf_{q \in [\underline{\tau}, \bar{\tau}]} \lambda_{\min}\{\mathbb{E}(X(\tau)X(\tau)^T|Q = q)\}P(E)\|\alpha - \alpha_0\|_2^2.
\end{aligned}
$$

This result combined with (C.2) implies that

$$X^T \delta_0 1\{\tau < Q \leq \tau_0\} = 0 \text{ a.s.}$$

This also implies that

$$
\begin{aligned}
0 &= \mathbb{E}[(X^T\delta_0)^2 1\{\tau < Q \leq \tau_0\}] \\
&= \mathbb{E}\{1\{\tau < Q \leq \tau_0\}\mathbb{E}((X^T\delta_0)^2|Q)\} \\
&\geq \inf_{q \in [\underline{\tau}, \bar{\tau}]} \mathbb{E}((X^T\delta_0)^2|Q = q)P(\tau < Q \leq \tau_0).
\end{aligned}
$$

Since it is assumed that $\inf_{q \in [\underline{\tau}, \bar{\tau}]} \mathbb{E}((X^T\delta_0)^2|Q = q) > 0$, thus $P(\tau < Q \leq \tau_0) = 0$. However, we also assume that $P(\tau < Q \leq \tau_0) > 0$ if $\tau < \tau_0$. Hence we must have $\tau = \tau_0$.

Now consider the other case, that is $\tau < \tau_0$. In this case, we have that

$$\mathcal{D}(\alpha, \tau) = X^T(\beta - \beta_0) + X^T(\delta - \delta_0)1\{Q > \tau\} + X^T\delta_0 1\{\tau_0 < Q \leq \tau\}. \tag{C.4}$$

44

Hence, in this case, modifying the definition of E in (C.3) to be

$$E \equiv \{1\{\tau_0 < Q \le \tau\} = 0\} = \{Q \notin (\tau_0, \tau]\}.$$

and proceeding the arguments identical to those above gives the desired result. ∎

C.2 Improved Risk Consistency

The following theorem shows that an improved rate of convergence is possible for the excess risk by taking the second and third steps of estimation. Recall that

$$\omega_n \propto \sqrt{\frac{\log(p \vee n)}{n}} \ .$$

Theorem C.2 (Improved Risk Consistency). *Let Assumption 1 hold. In addition, assume that $|\widehat{\tau} - \tau_0| = O_P(n^{-1})$ when $\delta_0 \ne 0$. Then, whether $\delta_0 = 0$ or not,*

$$R(\widehat{\alpha}, \widehat{\tau}) = O_P(\omega_n s).$$

The proof of this theorem is given in Appendix E.3. For the sake of not introducing additional assumptions in this section, we have assumed in Theorem C.2 that $|\widehat{\tau} - \tau_0| = O_P(n^{-1})$ when τ_0 is identifiable. Its formal statement is given by Theorem 4.3 in Section 4.

Remark C.1. As in Theorem 3.1, the risk consistency part of Theorem C.2 holds whether or not $\delta_0 = 0$. We obtain the improved rate of convergence in probability for the excess risk by combining the fact that our objective function is convex with respect to α given each τ with the second-step estimation results that (i) if $\delta \ne 0$, then $\widehat{\tau}$ is within a shrinking local neighborhood of τ_0, and (ii) when $\delta_0 = 0$, $\widehat{\tau}$ does not affect the excess risk in the sense that $R(\alpha_0, \tau) = 0$ for all $\tau \in \mathcal{T}$.

D   Regularity conditions on the general loss function

Let Y be a scalar variable of outcome and X be a vector of p-dimensional observed characteristics. Suppose there is an observable scalar variable Q such that the conditional distribution of Y or some feature of that (given X) depends on:

$$X^T \beta_0 1\{Q \le \tau_0\} + X^T \theta_0 1\{Q > \tau_0\} = X^T \beta_0 + X^T \delta_0 1\{Q > \tau_0\},$$

where $\delta_0 = \theta_0 - \beta_0$. Let $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ be a loss function under consideration, whose analytical form is clear in specific models. Suppose the true parameters are defined as the minimizer of the expected loss:

$$(\beta_0, \delta_0, \tau_0) \equiv \operatorname*{argmin}_{(\beta,\delta)\in\mathcal{A}, \tau\in\mathcal{T}} \mathbb{E}\left[\rho(Y, X^T\beta + X^T\delta 1\{Q > \tau\})\right], \tag{D.1}$$

where $\mathcal{A}$ and $\mathcal{T}$ denote the parameter spaces for (β_0, δ_0) and τ_0. Here β represents the components of "baseline parameters", while δ represents the structural changes; τ is the change point value where the structural changes occur, if any. By construction, τ_0 is not unique when $\delta_0 = 0$. For each $(\beta, \delta) \in \mathcal{A}$ and $\tau \in \mathcal{T}$, define $2p \times 1$ vectors:

$$\alpha \equiv (\beta^T, \delta^T)^T, \quad X(\tau) \equiv (X^T, X^T 1\{Q > \tau\})^T.$$

Then $X^T\beta + X^T\delta 1\{Q > \tau\} = X(\tau)^T\alpha$, and by letting $\alpha_0 \equiv (\beta_0^T, \delta_0^T)^T$, we can write (D.1) more compactly as:

$$(\alpha_0, \tau_0) = \operatorname*{argmin}_{\alpha\in\mathcal{A}, \tau\in\mathcal{T}} \mathbb{E}\left[\rho(Y, X(\tau)^T\alpha)\right]. \tag{D.2}$$

In quantile regression models, for a given quantile $\gamma \in (0, 1)$, recall that

$$\rho(t_1, t_2) = (t_1 - t_2)(\gamma - 1\{t_1 - t_2 \le 0\}).$$

D.1   When $\delta_0 \neq 0$ and τ_0 is identified

For a constant $\eta > 0$, define

$$r_1(\eta) \equiv \sup_r \Big\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \leq \tau_0\right\}\right)$$
$$\geq \eta\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}] \text{ for all } \beta \in \mathcal{B}(\beta_0, r) \Big\}$$

and

$$r_2(\eta) \equiv \sup_r \Big\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{Q > \tau_0\right\}\right)$$
$$\geq \eta\mathbb{E}[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}] \text{ for all } \theta \in \mathcal{G}(\theta_0, r) \Big\},$$

where $\mathcal{B}(\beta_0, r)$ and $\mathcal{G}(\theta_0, r)$ are defined in (B.3). Note that $r_1(\eta)$ and $r_2(\eta)$ are the maximal radii over which the excess risk can be bounded below by the quadratic loss on $\{Q \leq \tau_0\}$ and $\{Q > \tau_0\}$, respectively.

Assumption 7.   *(i) Let $\mathcal{Y}$ denote the support of Y. There is a Liptschitz constant $L > 0$ such that for all $y \in \mathcal{Y}$, $\rho(y, \cdot)$ is convex, and*

$$|\rho(y, t_1) - \rho(y, t_2)| \leq L|t_1 - t_2|, \forall t_1, t_2 \in \mathbb{R}.$$

(ii) For all $\alpha \in \mathcal{A}$, almost surely,

$$\mathbb{E}\left[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q\right] \geq 0.$$

(iii) There exist constants $\eta^* > 0$ and $r^* > 0$ such that $r_1(\eta^*) \geq r^*$ and $r_2(\eta^*) \geq r^*$.

(iv) There is a constant $c_0 > 0$ such that for all $\tau \in \mathcal{T}_0$,

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right)\mathbb{1}\left\{\tau < Q \leq \tau_0\right\}\right] \geq c_0\mathbb{E}\left[\left(X^T\left(\beta_0 - \theta_0\right)\right)^2\mathbb{1}\left\{\tau < Q \leq \tau_0\right\}\right],$$

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right)\mathbb{1}\left\{\tau_0 < Q \leq \tau\right\}\right] \geq c_0\mathbb{E}\left[\left(X^T\left(\beta_0 - \theta_0\right)\right)^2\mathbb{1}\left\{\tau_0 < Q \leq \tau\right\}\right].$$

We focus on a convex Lipchitz loss function, which is assumed in condition (i). It might be possible to weaken the convexity to a "restricted strong convexity condition" as in Loh and Wainwright (2013). For simplicity, we focus on the case of a convex loss, which is satisfied for quantile regression. However, unlike the framework of M-estimation in Negahban et al. (2012) and Loh and Wainwright (2013), we do allow $\rho(t_1, t_2)$ to be non-differentiable, which admits the quantile regression model as a special case.

Condition (iii) requires that the excess risk can be bounded below by a quadratic function locally when τ is fixed at τ_0, while condition (iv) is an analogous condition when α is fixed at α_0. conditions (iii) and (iv), combined with the convexity of $\rho(Y, \cdot)$, helps us derive the rates of convergence (in the ℓ_1 norm) of the Lasso estimators of (α_0, τ_0). Furthermore, these two conditions separate the conditions for α and τ, making them easier to interpret and verify.

Remark D.1. Condition (iii) of Assumption 7 is similar to *the restricted nonlinear impact (RNI)* condition of Belloni and Chernozhukov (2011). One may consider an alternative formulation as in van de Geer (2008) and Bühlmann and van de Geer (2011) (Chapter 6), which is known as the *margin condition*. But the margin condition needs to be adjusted to account for structural changes as in condition (iv). It would be an interesting future research topic to develop a general theory of high-dimensional M-estimation with an unknown sparsity-structural-change with general margin conditions.

Remark D.2. Assumptions 7 (iv) and 5 (iii) together imply that for all $\tau \in \mathcal{T}_0$, there exists

a constant $c_0 > 0$ such that

$$\Delta_1(\tau) \equiv \mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau < Q \le \tau_0\right\}\right] \ge c_0^2 \mathbb{P}\left[\tau < Q \le \tau_0\right],$$

$$\Delta_2(\tau) \equiv \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \le \tau\right\}\right] \ge c_0^2 \mathbb{P}\left[\tau_0 < Q \le \tau\right]. \tag{D.3}$$

Note that Assumption 7 (ii) implies that $\Delta_1(\tau)$ is monotonely non-increasing when $\tau < \tau_0$, and $\Delta_2(\tau)$ is monotonely non-decreasing when $\tau > \tau_0$, respectively. Therefore, Assumptions 7 (ii), 7 (iv) and 5 (iii) all together imply that (D.3) holds for all τ in the $\mathcal{T}$, not just in the $\mathcal{T}_0$ since $\mathcal{T}$ is compact. Equation (D.3) plays an important role in achieving a super-efficient convergence rate for τ_0, since it states the presence of a kink in the expected loss and that of a jump in the loss function at τ_0.

We now move to the set of assumptions that are useful to deal with the Step 3b estimator. Define

$$m_j(\tau, \alpha) \equiv \frac{\partial \mathbb{E}[\rho(Y, X(\tau)^T\alpha)]}{\partial \alpha_j}, \quad m(\tau, \alpha) \equiv (m_1(\tau, \alpha), ..., m_{2p}(\tau, \alpha))^T.$$

Also, let $m_J(\tau, \alpha) \equiv (m_j(\tau, \alpha) : j \in J(\alpha_0))$.

Assumption 8. $\mathbb{E}[\rho(Y, X(\tau)^T\alpha)]$ *is three times continuously differentiable with respect to* α*, and there are constants* $c_1, c_2, L > 0$ *and a neighborhood* $\mathcal{T}_0$ *of* τ_0 *such that the following conditions hold: for all large* n *and all* $\tau \in \mathcal{T}_0$*,*

(i) there is $M_n > 0$*, which may depend on the sample size* n*, such that*

$$\max_{j \le 2p} |m_j(\tau, \alpha_0) - m_j(\tau_0, \alpha_0)| < M_n|\tau - \tau_0|;$$

(ii) there is $r > 0$ *such that for all* $\beta \in \mathcal{B}(\beta_0, r)$*,* $\theta \in \mathcal{G}(\theta_0, r)$*,* $\alpha = (\beta^T, \theta^T - \beta^T)^T$ *satisfies:*

$$\max_{j \le 2p} \sup_{\tau \in \mathcal{T}_0} |m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| < L |\alpha - \alpha_0|_1;$$

(iii) α_0 is in the interior of the parameter space $\mathcal{A}$, and

$$\inf_{\tau \in \mathcal{T}_0} \lambda_{\min} \left(\frac{\partial^2 \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_{0J})]}{\partial \alpha_J \partial \alpha_J^T} \right) > c_1,$$

$$\sup_{|\alpha_J - \alpha_{0J}|_1 < c_2, \, \tau \in \mathcal{T}_0} \max_{i,j,k \in J} \left| \frac{\partial^3 \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_J)]}{\partial \alpha_i \partial \alpha_j \partial \alpha_k} \right| < L.$$

The score-condition in the population level is expressed by $m(\tau_0, \alpha_0) = 0$ since α_0 is in the interior of $\mathcal{A}$ by condition (iii). Conditions (i) and (ii) regulate the continuity of the score $m(\tau, \alpha)$, and condition (iii) assumes the higher-order differentiability of the expectation of the loss function. Condition (i) requires the Lipschitz continuity of the score function with respect to the threshold. The Lipschitz constant may grow with n, since it is assumed uniformly over $j \leq 2p$. In many examples, M_n in fact grows slowly; as a result, it does not affect the asymptotic behavior of $\widetilde{\alpha}$. For quantile regression models, we will show that $M_n = Cs^{1/2}$ for some constant $C > 0$. Condition (ii) requires the local equicontinuity at α_0 in the ℓ_1 norm of the class

$$\{m_j(\tau, \alpha) : \tau \in \mathcal{T}_0, j \leq 2p\}.$$

We now establish that Assumptions 7 and 8 are satisfied for quantile regression models.

Lemma D.1. *Suppose that Assumptions 1 and 2 hold. Then Assumptions 7 and 8 are satisfied by the loss function for the quantile regression model, with $M_n = Cs^{1/2}$ for some constant $C > 0$.*

D.1.1 Proof of Lemma D.1

Verification of Assumption 7 (i). It is straightforward to show that the loss function for quantile regression is convex and satisfies the Liptschitz condition. ∎

Verification of Assumption 7 (ii). Note that $\rho(Y, t) = h_\gamma(Y - t)$, where $h_\gamma(t) = t(\gamma - 1\{t \leq$

0$}$). By (B.3) of Belloni and Chernozhukov (2011),

$$h_\gamma(w - v) - h_\gamma(w) = -v(\gamma - 1\{w \le 0\}) + \int_0^v (1\{w \le z\} - 1\{w \le 0\})dz \qquad \text{(D.4)}$$

where $w = Y - X(\tau_0)^T \alpha_0$ and $v = X(\tau_0)^T(\alpha - \alpha_0)$. Note that

$$\mathbb{E}[v(\gamma - 1\{w \le 0\})|Q] = -\mathbb{E}[X(\tau_0)^T(\alpha - \alpha_0)(\gamma - 1\{U \le 0\})|Q] = 0,$$

since $\mathbb{P}(U \le 0|X, Q) = \gamma$. Let $F_{Y|X,Q}$ denote the CDF of the conditional distribution $Y|X, Q$.

Then

$$
\begin{aligned}
&\mathbb{E}\left[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q\right] \\
&= \mathbb{E}\left[\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} (1\{U \le z\} - 1\{U \le 0\})dz \Big| Q\right] \\
&= \mathbb{E}\left[\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} [F_{Y|X,Q}(X(\tau_0)^T\alpha_0 + z|X, Q) - F_{Y|X,Q}(X(\tau_0)^T\alpha_0|X, Q)]dz \Big| Q\right] \\
&\ge 0,
\end{aligned}
$$

where the last inequality follows immediately from the fact that $F_{Y|X,Q}(\cdot|X, Q)$ is the CDF.

Hence, we have verified Assumption 7 (ii). ∎

Verification of Assumption 7 (iii). Following the arguments analogous those used in (B.4) of Belloni and Chernozhukov (2011), the mean value expansion implies:

$$
\begin{aligned}
&\mathbb{E}\left[\rho(Y, X(\tau_0)^T\alpha) - \rho(Y, X(\tau_0)^T\alpha_0)|Q\right] \\
&= \mathbb{E}\left\{\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} \left[z f_{Y|X,Q}(X(\tau_0)^T\alpha_0|X, Q) + \frac{z^2}{2}\tilde{f}_{Y|X,Q}(X(\tau_0)^T\alpha_0 + t|X, Q)\right] dz \Big| Q\right\} \\
&= \frac{1}{2}(\alpha - \alpha_0)^T \mathbb{E}\left[X(\tau_0)X(\tau_0)^T f_{Y|X,Q}(X(\tau_0)^T\alpha_0|X, Q)|Q\right](\alpha - \alpha_0) \\
&\quad + \mathbb{E}\left\{\int_0^{X(\tau_0)^T(\alpha-\alpha_0)} \frac{z^2}{2}\tilde{f}_{Y|X,Q}(X(\tau_0)^T\alpha_0 + t|X, Q)dz \Big| Q\right\}
\end{aligned}
$$

51

for some intermediate value t between 0 and z. By condition (ii) of Assumption 2,

$$|\tilde{f}_{Y|X,Q}(X(\tau_0)^T\alpha_0 + t|X, Q)| \leq C_1 \quad \text{and} \quad f_{Y|X,Q}(X(\tau_0)^T\alpha_0|X, Q) \geq C_2.$$

Hence, taking the expectation on $\{Q \leq \tau_0\}$ gives

$$\mathbb{E}\left[\rho(Y, X^T\beta) - \rho(Y, X^T\beta_0)1\{Q \leq \tau_0\}\right]$$
$$\geq \frac{C_2}{2}\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}] - \frac{C_1}{6}\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \leq \tau_0\}]$$
$$\geq \frac{C_2}{4}\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \leq \tau_0\}],$$

where the last inequality follows from

$$\frac{C_2}{4}\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \leq \tau_0\}] \geq \frac{C_1}{6}\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \leq \tau_0\}]. \tag{D.5}$$

To see why (D.5) holds, note that by (B.5), for any nonzero $\beta \in \mathcal{B}(\beta_0, r_{QR}^*)$,

$$\frac{\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \leq \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \leq \tau_0\}]} \geq r_{QR}^* \frac{2C_1}{3C_2} \geq \frac{2C_1}{3C_2}\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \leq \tau_0\}]^{1/2},$$

which proves (D.5) immediately. Thus, we have shown that Assumption 7 (iii) holds for $r_1(\eta)$ with $\eta^* = C_2/4$ and $r^* = r_{QR}^*$ defined in (B.5) in Assumption 2. The case for $r_2(\eta)$ is similar and hence is omitted. ∎

Verification of Assumption 7 (iv). We again start from (D.4) but with different choices of (w, v) such that $w = Y - X(\tau_0)^T\alpha_0$ and $v = X^T\delta_0[1\{Q \leq \tau_0\} - 1\{Q > \tau_0\}]$. Then arguments

similar to those used in verifying Assumptions 7 (ii)-(iii) yield that for $\tau < \tau_0$,

$$\mathbb{E}\left[\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)|Q = \tau\right] \tag{D.6}$$

$$= \mathbb{E}\left\{\int_0^{X^T\delta_0} z f_{Y|X,Q}(X^T\beta_0 + t|X, Q)dz\Big|Q = \tau\right\} \tag{D.7}$$

$$\geq \mathbb{E}\left\{\int_0^{\widetilde{\varepsilon}(X^T\delta_0)} z f_{Y|X,Q}(X^T\beta_0 + t|X, Q)dz\Big|Q = \tau\right\} \tag{D.8}$$

$$\geq \frac{\widetilde{\varepsilon}^2 C_3}{2}\mathbb{E}\left[(X^T\delta_0)^2|Q = \tau\right], \tag{D.9}$$

where t is an intermediate value t between 0 and z. Thus, we have that

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right)\mathbf{1}\left\{\tau < Q \leq \tau_0\right\}\right] \geq \frac{\widetilde{\varepsilon}^2 C_3}{2}\mathbb{E}\left[(X^T(\beta_0 - \theta_0))^2\mathbf{1}\left\{\tau < Q \leq \tau_0\right\}\right].$$

The case that $\tau > \tau_0$ is similar. $\blacksquare$

Verification of Assumption 8. Note that

$$m_j(\tau, \alpha) = \mathbb{E}[X_j(\tau)(\mathbf{1}\{Y - X(\tau)^T\alpha \leq 0\} - \gamma)].$$

Hence, $m_j(\tau_0, \alpha_0) = 0$, for all $j \leq 2p$. For condition (i) of Assumption 8, for all $j \leq 2p$,

$$|m_j(\tau, \alpha_0) - m_j(\tau_0, \alpha_0)|$$

$$= |\mathbb{E}X_j(\tau)[1\{Y \leq X(\tau)^T\alpha_0\} - 1\{Y \leq X(\tau_0)^T\alpha_0\}]|$$

$$= |\mathbb{E}X_j(\tau)[\mathbb{P}(Y \leq X(\tau)^T\alpha_0|X, Q) - \mathbb{P}(Y \leq X(\tau_0)^T\alpha_0|X, Q)]|$$

$$\leq C\mathbb{E}|X_j(\tau)||(X(\tau) - X(\tau_0))^T\alpha_0|$$

$$= C\mathbb{E}|X_j(\tau)||X^T\delta_0(1\{Q > \tau\} - 1\{Q > \tau_0\})|$$

$$\leq C\mathbb{E}|X_j(\tau)||X^T\delta_0|(1\{\tau < Q < \tau_0\} + 1\{\tau_0 < Q < \tau\})$$

$$\leq C(\mathbb{P}(\tau_0 < Q < \tau) + \mathbb{P}(\tau < Q < \tau_0)) \sup_{\tau, \tau' \in \mathcal{T}_0} \mathbb{E}(|X_j(\tau)X^T\delta_0||Q = \tau')$$

$$\leq C(\mathbb{P}(\tau_0 < Q < \tau) + \mathbb{P}(\tau < Q < \tau_0)) \sup_{\tau, \tau' \in \mathcal{T}_0} [\mathbb{E}(|X_j(\tau)|^2||Q = \tau')]^{1/2}[\mathbb{E}(|X^T\delta_0|^2|Q = \tau')]^{1/2}$$

$$\leq CM_2K_2|\delta_0|_2|\tau_0 - \tau|$$

for some constant C, where the last inequality follows from conditions (ii), (iii) and (v) of Assumption 1. Therefore, we have verified condition (i) of Assumption 8 with $M_n = CM_2K_2|\delta_0|_2$.

We now verify condition (ii) of Assumption 8. For all j and τ in a neighborhood of τ_0,

$$|m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| = |\mathbb{E}X_j(\tau)(1\{Y \leq X(\tau)^T\alpha\} - 1\{Y \leq X(\tau)^T\alpha_0\})|$$

$$= |\mathbb{E}X_j(\tau)(\mathbb{P}(Y \leq X(\tau)^T\alpha|X, Q) - \mathbb{P}(Y \leq X(\tau)^T\alpha_0|X, Q))|$$

$$\leq C\mathbb{E}|X_j(\tau)||X(\tau)^T(\alpha - \alpha_0)| \leq C|\alpha - \alpha_0|_1 \max_{j \leq 2p, i \leq 2p} \mathbb{E}|X_j(\tau)X_i(\tau)|,$$

which implies the result immediately in view of Assumption 1. Finally, it is straightforward to verify condition (iii) using Assumption 2 (iii). ∎

D.2   When $\delta_0 = 0$

We now consider the case when $\delta_0 = 0$. In this case, τ_0 is not identifiable, and there is actually no structural change in the sparsity. If α_0 is in the interior of $\mathcal{A}$, then $m(\tau, \alpha_0) = 0$ for all $\tau \in \mathcal{T}$.

For a constant $\eta > 0$, define

$$\tilde{r}_1(\eta) \equiv \sup_r \left\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right) \right] 1\left\{Q \le \tau\right\} \right) \right.$$
$$\left. \ge \eta\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau\}] \text{ for all } \beta \in \tilde{\mathcal{B}}(\beta_0, r, \tau) \text{ and for all } \tau \in \mathcal{T} \right\}$$

and

$$\tilde{r}_2(\eta) \equiv \sup_r \left\{ r : \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right) \right] 1\left\{Q > \tau\right\} \right) \right.$$
$$\left. \ge \eta\mathbb{E}[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}] \text{ for all } \theta \in \tilde{\mathcal{G}}(\beta_0, r, \tau) \text{ and for all } \tau \in \mathcal{T} \right\},$$

where $\tilde{\mathcal{B}}(\beta_0, r, \tau)$ and $\tilde{\mathcal{G}}(\beta_0, r, \tau)$ are defined in (B.4).

Assumption 9.   *(i) Let $\mathcal{Y}$ denote the support of Y. There is a Liptschitz constant $L > 0$ such that for all $y \in \mathcal{Y}$, $\rho(y, \cdot)$ is convex, and*

$$|\rho(y, t_1) - \rho(y, t_2)| \le L|t_1 - t_2|, \forall t_1, t_2 \in \mathbb{R}.$$

(ii) For all $\alpha \in \mathcal{A}$ and for all $\tau \in \mathcal{T}$, almost surely,

$$\mathbb{E}[\rho(Y, X(\tau)^T\alpha) - \rho(Y, X^T\beta_0)|Q] \ge 0,$$

(iii) There exist constants $\eta^* > 0$ and $r^* > 0$ such that $\tilde{r}_1(\eta^*) \ge r^*$ and $\tilde{r}_2(\eta^*) \ge r^*$.

(iv) $\mathbb{E}[\rho(Y, X(\tau)^T\alpha)]$ is three times differentiable with respect to α, and there are universal constants $r > 0$ and $L > 0$ such that for all $\beta \in \tilde{\mathcal{B}}(\beta_0, r, \tau)$, $\theta \in \tilde{\mathcal{G}}(\beta_0, r, \tau)$, $\alpha =$

$(\beta^T, \theta^T - \beta^T)^T$ satisfies:

$$\max_{j \le 2p} |m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| < L \, |\alpha - \alpha_0|_1 \, .$$

for all large n and for all $\tau \in \mathcal{T}$.

(v) α_0 is in the interior of the parameter space $\mathcal{A}$, and there are constants c_1 an $c_2 > 0$ such that

$$\lambda_{\min} \left(\frac{\partial^2 \mathbb{E}[\rho(Y, X_{J(\beta_0)}^T \beta_{0J})]}{\partial \beta_J \partial \beta_J^T} \right) > c_1,$$

$$\sup_{|\alpha_J - \alpha_{0J}|_1 < c_2,} \max_{i,j,k \in J(\beta_0)} \left| \frac{\partial^3 \mathbb{E}[\rho(Y, X_{J(\beta_0)}^T \beta_J)]}{\partial \beta_i \partial \beta_j \partial \beta_k} \right| < L.$$

As in Lemma D.1, we now establish that Assumption 9 is satisfied for quantile regression models when $\delta_0 = 0$.

Lemma D.2. *Suppose that Assumptions 1 and 2 hold. Then Assumption 9 is satisfied.*

D.2.1 Proof of Lemma D.2

Verification of Assumption 9 (i). This is the same as the verification of Assumption 7 (i). ∎

Verification of Assumption 9 (ii). This can be verified exactly as in verification of Assumption 7 (ii) with $\alpha_0 = \beta_0$ now. ∎

Verification of Assumption 9 (iii). By the arguments identical to those used to verify Assumption 7 (iii), we have that

$$\mathbb{E}\left[\rho(Y, X^T \beta) - \rho(Y, X^T \beta_0) 1\{Q \le \tau\} \right]$$
$$\ge \frac{C_2}{2} \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau\}] - \frac{C_1}{6} \mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \le \tau\}]$$
$$\ge \frac{C_2}{4} \mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \le \tau\}],$$

where the last inequality follows from (B.7). This proves the case for $\tilde{r}_1(\eta)$. The case for $\tilde{r}_2(\eta)$ is similar and hence is omitted. ∎

Verification of Assumptions 9 (iv) and (v). They can be verified similarly as in verification of Assumption 8 in the proof of Lemma Lemma D.1. For all j and $\tau \in \mathcal{T}$,

$$
\begin{aligned}
|m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| &= |\mathbb{E}X_j(\tau)(1\{Y \leq X(\tau)^T\alpha\} - 1\{Y \leq X(\tau)^T\alpha_0\})| \\
&= |\mathbb{E}X_j(\tau)(\mathbb{P}(Y \leq X(\tau)^T\alpha|X, Q) - \mathbb{P}(Y \leq X(\tau)^T\alpha_0|X, Q))| \\
&\leq C\mathbb{E}|X_j(\tau)||X(\tau)^T(\alpha - \alpha_0)| \leq C|\alpha - \alpha_0|_1 \max_{j \leq 2p, i \leq 2p} \mathbb{E}|X_j(\tau)X_i(\tau)|,
\end{aligned}
$$

which implies condition 9 (iv) in view of Assumption 1. It is also straightforward to verify condition 9 (v) using Assumption 2 (iii). ∎

E   Proofs of Theorems

Throughout the proofs, we define

$$
\nu_n(\alpha, \tau) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[\rho\left(Y_i, X_i(\tau)^T \alpha\right) - \mathbb{E}\rho\left(Y, X(\tau)^T \alpha\right) \right].
$$

Without loss of generality let $\nu_n(\alpha_J, \tau) = n^{-1} \sum_{i=1}^{n} \left[\rho\left(Y_i, X_{iJ}(\tau)^T \alpha_J\right) - \mathbb{E}\rho\left(Y, X_J(\tau)^T \alpha_J\right) \right]$.

In this section, we suppose that Assumptions 7 and 8 hold when $\delta_0 \neq 0$ and that Assumption 9 holds when $\delta_0 = 0$, respectively.

E.1   Useful Lemmas

For the positive constant K_1 in Assumption 1 (i), define

$$
c_{np} \equiv \sqrt{\frac{2\log(4np)}{n}} + \frac{K_1 \log(4np)}{n}.
$$

Let $\lceil x \rceil$ denote the smallest integer greater than or equal to a real number x. The following lemma bounds $\nu_n(\alpha, \tau)$.

Lemma E.1. *For any positive sequences m_{1n} and m_{2n}, and any $\widetilde{\delta} \in (0, 1)$, there are constants L_1, L_2 and $L_3 > 0$ such that for $a_n = L_1 c_{np} \widetilde{\delta}^{-1}$, $b_n = L_2 c_{np} \lceil \log_2(m_{2n}/m_{1n}) \rceil \widetilde{\delta}^{-1}$, and $c_n = L_3 n^{-1/2} \widetilde{\delta}^{-1}$,*

$$\mathbb{P} \left\{ \sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} |\nu_n(\alpha, \tau) - \nu_n(\alpha_0, \tau)| \geq a_n m_{1n} \right\} \leq \widetilde{\delta}, \tag{E.1}$$

$$\mathbb{P} \left\{ \sup_{\tau \in \mathcal{T}} \sup_{m_{1n} \leq |\alpha - \alpha_0|_1 \leq m_{2n}} \frac{|\nu_n(\alpha, \tau) - \nu_n(\alpha_0, \tau)|}{|\alpha - \alpha_0|_1} \geq b_n \right\} \leq \widetilde{\delta}, \tag{E.2}$$

and for any $\eta > 0$ and $\mathcal{T}_\eta = \{\tau \in \mathcal{T} : |\tau - \tau_0| \leq \eta\}$,

$$\mathbb{P} \left\{ \sup_{\tau \in \mathcal{T}_\eta} |\nu_n(\alpha_0, \tau) - \nu_n(\alpha_0, \tau_0)| \geq c_n |\delta_0|_2 \sqrt{\eta} \right\} \leq \widetilde{\delta}. \tag{E.3}$$

Proof of (E.1): Let $\epsilon_1, ..., \epsilon_n$ denote a Rademacher sequence, independent of $\{Y_i, X_i, Q_i\}_{i \leq n}$. By the symmetrization theorem (see, for example, Theorem 14.3 of Bühlmann and van de Geer (2011)) and then by the contraction theorem (see, for example, Theorem 14.4 of Bühlmann and van de Geer (2011)),

$$\mathbb{E} \left(\sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} |\nu_n(\alpha, \tau) - \nu_n(\alpha_0, \tau)| \right)$$

$$\leq 2 \mathbb{E} \left(\sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \left[\rho\left(Y_i, X_i(\tau)^T \alpha\right) - \rho\left(Y_i, X_i(\tau)^T \alpha_0\right) \right] \right| \right)$$

$$\leq 4L \mathbb{E} \left(\sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i X_i(\tau)^T (\alpha - \alpha_0) \right| \right).$$

Note that

$$
\begin{aligned}
\sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} & \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i X_i(\tau)^T (\alpha - \alpha_0) \right| \\
&= \sup_{\tau \in \mathcal{T}} \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} \left| \sum_{j=1}^{2p} (\alpha_j - \alpha_{0j}) \frac{1}{n} \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right| \\
&\leq \sup_{|\alpha - \alpha_0|_1 \leq m_{1n}} \sum_{j=1}^{2p} |\alpha_j - \alpha_{0j}| \sup_{\tau \in \mathcal{T}} \max_{j \leq 2p} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right| \\
&\leq m_{1n} \sup_{\tau \in \mathcal{T}} \max_{j \leq 2p} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right|.
\end{aligned}
\tag{E.4}
$$

For all $\tilde{L} > K_1$,

$$
\begin{aligned}
\mathbb{E} \left(\sup_{\tau \in \mathcal{T}} \max_{j \leq 2p} \left| \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right| \right) &\leq_{(1)} \tilde{L} \log \mathbb{E} \left[\exp \left(\tilde{L}^{-1} \sup_{\tau \in \mathcal{T}} \max_{j \leq 2p} \left| \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right| \right) \right] \\
&\leq_{(2)} \tilde{L} \log \mathbb{E} \left[\exp \left(\tilde{L}^{-1} \max_{\tau \in \{Q_1, \ldots, Q_n\}} \max_{j \leq 2p} \left| \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right| \right) \right] \\
&\leq_{(3)} \tilde{L} \log \left[4np \exp \left(\frac{n}{2(\tilde{L}^2 - \tilde{L} K_1)} \right) \right],
\end{aligned}
$$

where inequality (1) follows from Jensen's inequality, inequality (2) comes from the fact that $X_{ij}(\tau)$ is a step function with jump points on $\mathcal{T} \cap \{Q_1, \ldots, Q_n\}$, and inequality (3) is by Bernstein's inequality for the exponential moment of an average (see, for example, Lemma 14.8 of Bühlmann and van de Geer (2011)), combined with the simple inequalities that $\exp(|x|) \leq \exp(x) + \exp(-x)$ and that $\exp(\max_{1 \leq j \leq J} x_j) \leq \sum_{j=1}^{J} \exp(x_j)$. Then it follows that

$$
\mathbb{E} \left(\sup_{\tau \in \mathcal{T}} \max_{j \leq 2p} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i X_{ij}(\tau) \right| \right) \leq \frac{\tilde{L} \log(4np)}{n} + \frac{1}{2(\tilde{L} - K_1)} = c_{np},
\tag{E.5}
$$

where the last equality follows by taking $\tilde{L} = K_1 + \sqrt{n/[2 \log(4np)]}$. Thus, by Markov's

inequality,

$$\mathbb{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{|\alpha-\alpha_0|_1\leq m_{1n}}|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|>a_n m_{1n}\right\}\leq(a_n m_{1n})^{-1}4Lm_{1n}c_{np}=\widetilde{\delta},$$

where the last equality follows by setting $L_1=4L$.

Proof of (E.2): Recall that $\epsilon_1,...,\epsilon_n$ is a Rademacher sequence, independent of $\{Y_i,X_i,Q_i\}_{i\leq n}$.
Note that

$$\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\leq|\alpha-\alpha_0|_1\leq m_{2n}}\frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1}\right)$$

$$\leq_{(1)}2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\leq|\alpha-\alpha_0|_1\leq m_{2n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\frac{\rho\left(Y_i,X_i(\tau)^T\alpha\right)-\rho\left(Y_i,X_i(\tau)^T\alpha_0\right)}{|\alpha-\alpha_0|_1}\right|\right)$$

$$\leq_{(2)}2\sum_{j=1}^k\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{2^{j-1}m_{1n}\leq|\alpha-\alpha_0|_1\leq 2^j m_{1n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\frac{\rho\left(Y_i,X_i(\tau)^T\alpha\right)-\rho\left(Y_i,X_i(\tau)^T\alpha_0\right)}{2^{j-1}m_{1n}}\right|\right)$$

$$\leq_{(3)}4L\sum_{j=1}^k\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\sup_{2^{j-1}m_{1n}\leq|\alpha-\alpha_0|_1\leq 2^j m_{1n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\frac{X_i(\tau)^T(\alpha-\alpha_0)}{2^{j-1}m_{1n}}\right|\right),$$

where inequality (1) is by the symmetrization theorem, inequality (2) holds for some $k\equiv\lceil\log_2(m_{2n}/m_{1n})\rceil$, and inequality (3) follows from the contraction theorem.

Next, the identical arguments showing (E.4) yield

$$\sup_{2^{j-1}m_{1n}\leq|\alpha-\alpha_0|_1\leq 2^j m_{1n}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\frac{X_i(\tau)^T(\alpha-\alpha_0)}{2^{j-1}m_{1n}}\right|\leq 2\max_{j\leq 2p}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i X_{ij}(\tau)\right|$$

uniformly in $\tau\in\mathcal{T}$. Then, as in the proof of (E.1), Bernstein's and Markov's inequalities imply that

$$\mathbb{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{m_{1n}\leq|\alpha-\alpha_0|_1\leq m_{2n}}\frac{|\nu_n(\alpha,\tau)-\nu_n(\alpha_0,\tau)|}{|\alpha-\alpha_0|_1}>b_n\right\}\leq b_n^{-1}8Lkc_{np}=\widetilde{\delta},$$

where the last equality follows by setting $L_2=8L$.

Proof of (E.3): As above, by the symmetrization and contraction theorems, we have that

$$
\mathbb{E} \left(\sup_{\tau \in \mathcal{T}_\eta} |\nu_n (\alpha_0, \tau) - \nu_n (\alpha_0, \tau_0)| \right)
$$

$$
\leq \quad 2\mathbb{E} \left(\sup_{\tau \in \mathcal{T}_\eta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \left[\rho \left(Y_i, X_i (\tau)^T \alpha_0 \right) - \rho \left(Y_i, X_i (\tau_0)^T \alpha_0 \right) \right| \right] \right)
$$

$$
\leq \quad 4L\mathbb{E} \left(\sup_{\tau \in \mathcal{T}_\eta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i X_i^T \delta_0 \left(1 \{ Q_i > \tau \} - 1 \{ Q_i > \tau_0 \} \right) \right| \right)
$$

$$
\leq \quad \frac{4LC_1 (M_2 |\delta_0|_2^2 K_2 \eta)^{1/2}}{\sqrt{n}}
$$

for some constant $C_1 < \infty$, where the last inequality is due to Theorem 2.14.1 of van der Vaart and Wellner (1996) with M_2 in Assumption 1 (v) and K_2 in Assumption 1 (ii). Specifically, we apply the second inequality of this theorem to the class $\mathcal{F} = \{ f(\epsilon, X, Q, \tau) = \epsilon X^T \delta_0 (1\{Q > \tau\} - 1\{Q > \tau_0\}), \tau \in \mathcal{T}_\eta \}$. Note that $\mathcal{F}$ is a Vapnik-Cervonenkis class, which has a uniformly bounded entropy integral and thus $J(1, \mathcal{F})$ in their theorem is bounded, and that the L_2 norm of the envelope $|\epsilon_i X_i^T \delta_0| 1\{|Q_i - \tau_0| < \eta\}$ is proportional to the square root of the length of $\mathcal{T}_\eta$:

$$
(E|\epsilon_i X_i^T \delta_0|^2 1\{|Q_i - \tau_0| < \eta\})^{1/2} \leq (2M_2 |\delta_0|_2^2 K_2 \eta)^{1/2}.
$$

This implies the last inequality with C_1 being $\sqrt{2}$ times the entropy integral of the class $\mathcal{F}$. Then, by Markov's inequality, we obtain (E.3) with $L_3 = 4LC_1 (M_2 K_2)^{1/2}$.

E.2   Proof of Theorem 3.1

Define $D(\tau) = \text{diag}(D_j(\tau) : j \leq 2p)$; and also let $D_0 = D(\tau_0)$ and $\check{D} = D(\check{\tau})$. It follows from the definition of $(\check{\alpha}, \check{\tau})$ in (2.2) that

$$
\frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_i(\check{\tau})^T \check{\alpha}) + \kappa_n |\check{D}\check{\alpha}|_1 \leq \frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_i(\tau_0)^T \alpha_0) + \kappa_n |D_0 \alpha_0|_1. \tag{E.6}
$$

From (E.6) we obtain the following inequality

$$
\begin{aligned}
R(\breve{\alpha}, \breve{\tau}) \ &\leq \ [\nu_n(\alpha_0, \tau_0) - \nu_n(\breve{\alpha}, \breve{\tau})] + \kappa_n |D_0\alpha_0|_1 - \kappa_n|\breve{D}\breve{\alpha}|_1 \\
&= \ [\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})] + [\nu_n(\alpha_0, \tau_0) - \nu_n(\alpha_0, \breve{\tau})] \qquad \text{(E.7)} \\
&\quad + \kappa_n \left(|D_0\alpha_0|_1 - |\breve{D}\breve{\alpha}|_1 \right).
\end{aligned}
$$

Note that the second component $[\nu_n(\alpha_0, \tau_0) - \nu_n(\alpha_0, \breve{\tau})] = o_P\left[(s/n)^{1/2}\log n\right]$ due to (E.3) of Lemma E.1 with taking $\mathcal{T}_\eta = \mathcal{T}$ by choosing some sufficiently large $\eta > 0$. Thus, we focus on the other two terms in the following discussion. We consider two cases respectively: $|\breve{\alpha} - \alpha_0|_1 \leq |\alpha_0|_1$ and $|\breve{\alpha} - \alpha_0|_1 > |\alpha_0|_1$.

Suppose that $|\breve{\alpha} - \alpha_0|_1 \leq |\alpha_0|_1$. Then, $\left|\breve{D}\breve{\alpha}\right|_1 \leq \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 + \left|\breve{D}\alpha_0\right|_1 \leq 2\bar{D}|\alpha_0|_1$, and

$$
\left|\kappa_n \left(|D_0\alpha_0|_1 - |\breve{D}\breve{\alpha}|_1 \right)\right| \leq 3\kappa_n\bar{D}|\alpha_0|_1.
$$

Applying (E.1) in Lemma E.1 with $m_{1n} = |\alpha_0|_1$, we obtain

$$
|\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})| \leq a_n|\alpha_0|_1 \leq \kappa_n|\alpha_0|_1 \quad \text{w.p.a.1,}
$$

where the last inequality follows from the fact that $a_n \ll \kappa_n$ with κ_n satisfying (2.3). Thus, the theorem follows in this case.

Now assume that $|\breve{\alpha} - \alpha_0|_1 > |\alpha_0|_1$. In this case, apply (E.2) of Lemma E.1 with $m_{1n} = |\alpha_0|_1$ and $m_{2n} = 2M_1p$, where M_1 is defined in Assumption 1(iii), to obtain

$$
\frac{|\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})|}{|\breve{\alpha} - \alpha_0|_1} \leq b_n
$$

with probability arbitrarily close to one for small enough $\widetilde{\delta}$. Since $b_n \ll \underline{D}\kappa_n$, we have

$$
|\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})| \leq \kappa_n\underline{D}|\breve{\alpha} - \alpha_0|_1 \leq \kappa_n\left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 \quad \text{w.p.a.1.}
$$

Therefore,

$$R(\breve{\alpha}, \breve{\tau}) + o_P\left(n^{-1/2}\log n\right) \leq \kappa_n\left(|D_0\alpha_0|_1 - |\breve{D}\breve{\alpha}|_1\right) + \kappa_n\left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)\right|_1$$

$$\leq \kappa_n\left(|D_0\alpha_0|_1 - |\breve{D}\breve{\alpha}_J|_1\right) + \kappa_n\left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)_J\right|_1,$$

where the last inequality follows from the fact that $\breve{\alpha} - \alpha_0 = \breve{\alpha}_{J^C} + \left(\breve{\alpha} - \alpha_0\right)_J$. Thus, the theorem follows in this case as well.

E.3 Proof of Theorem C.2

Define

$$M^* \equiv 4\max_{\tau \in T_n}\left(R\left(\alpha_0, \tau\right) + 2\omega_n\bar{D}\left|\alpha_0\right|_1\right)/(\omega_n\underline{D}), \tag{E.8}$$

where $T_n \subset \mathcal{T}$ will be specified below. For each τ, define

$$\widehat{\alpha}(\tau) = \operatorname{argmin}_{\alpha \in \mathcal{A}}R_n(\alpha, \tau) + \omega_n\sum_{j=1}^{2p}D_j(\tau)|\alpha_j|. \tag{E.9}$$

It follows from the definition of $\widehat{\alpha}(\tau)$ in (E.9) that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\widehat{\alpha}(\tau)) + \omega_n|D(\tau)\widehat{\alpha}(\tau)|_1 \leq \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\alpha_0) + \omega_n|D(\tau)\alpha_0|_1. \tag{E.10}$$

Next, let

$$t\left(\tau\right) = \frac{M^*}{M^* + |\widehat{\alpha}\left(\tau\right) - \alpha_0|_1}$$

and $\bar{\alpha}\left(\tau\right) = t\left(\tau\right)\widehat{\alpha}\left(\tau\right) + \left(1 - t\left(\tau\right)\right)\alpha_0$. By construction, it follows that $|\bar{\alpha}\left(\tau\right) - \alpha_0|_1 \leq M^*$. And also note that

$$|\bar{\alpha}\left(\tau\right) - \alpha_0|_1 \leq M^*/2 \text{ implies } |\widehat{\alpha}\left(\tau\right) - \alpha_0|_1 \leq M^* \tag{E.11}$$

since $\bar{\alpha}\left(\tau\right) - \alpha_0 = t\left(\tau\right)\left(\widehat{\alpha}\left(\tau\right) - \alpha_0\right)$.

For each τ, (E.10) and the convexity of the following map

$$\alpha \mapsto \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\alpha) + \omega_n|D(\tau)\alpha|_1$$

implies that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\bar{\alpha}\left(\tau\right)) + \omega_n|D(\tau)\bar{\alpha}\left(\tau\right)|_1$$
$$\leq t\left(\tau\right)\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\widehat{\alpha}(\tau)) + \omega_n|D(\tau)\widehat{\alpha}(\tau)|_1\right]$$
$$+ \left[1 - t\left(\tau\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\alpha_0) + \omega_n|D(\tau)\alpha_0|_1\right]$$
$$\leq \left[\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\tau)^T\alpha_0) + \omega_n|D(\tau)\alpha_0|_1\right],$$

which in turn yields the following inequality

$$R(\bar{\alpha}(\tau), \tau) + \omega_n|D(\tau)\bar{\alpha}(\tau)|_1 \leq [\nu_n(\alpha_0, \tau) - \nu_n(\bar{\alpha}(\tau), \tau)] + R(\alpha_0, \tau) + \omega_n|D(\tau)\alpha_0|_1. \quad \text{(E.12)}$$

Furthermore, by the triangle inequality, (E.12) can be written as

$$R(\bar{\alpha}(\tau), \tau) + \omega_n\underline{D}|\bar{\alpha}(\tau) - \alpha_0|_1 \leq [\nu_n(\alpha_0, \tau) - \nu_n(\bar{\alpha}(\tau), \tau)] + R(\alpha_0, \tau) + 2\omega_n\overline{D}|\alpha_0|_1. \quad \text{(E.13)}$$

Now let $Z_M = \sup_{\tau \in T_n}\sup_{|\alpha - \alpha_0|\leq M}|\nu_n\left(\alpha, \tau\right) - \nu_n\left(\alpha_0, \tau\right)|$ for each $M > 0$. Then, by Lemma E.1, $Z_{M^*} = o_P\left(\omega_n M^*\right)$ by the simple fact that $\log(np) \leq 2\log(n \vee p)$. Thus, in view of the definition of M^* in (E.8), the following inequality holds w.p.a.1,

$$R(\bar{\alpha}(\tau), \tau) + \omega_n\underline{D}|\bar{\alpha}(\tau) - \alpha_0|_1 \leq \omega_n\underline{D}M^*/2 \quad \text{(E.14)}$$

uniformly in $\tau \in T_n$.

We can repeat the same arguments for $\widehat{\alpha}(\tau)$ instead of $\bar{\alpha}(\tau)$ due to (E.11) and (E.14), to obtain

$$R(\widehat{\alpha}(\tau), \tau) + \omega_n \underline{D} \, |\widehat{\alpha}(\tau) - \alpha_0|_1 \leq \omega_n \underline{D} M^* = O(\omega_n s), \text{ w.p.a.1,} \qquad \text{(E.15)}$$

uniformly in $\tau \in T_n$. It remains to show that there exists a set T_n such that $\widehat{\tau} \in T_n$ w.p.a.1 and the corresponding $M^* = O(s)$. We split the remaining part of the proof into two cases: $\delta_0 \neq 0$ and $\delta_0 = 0$.

(Case 1: $\delta_0 \neq 0$)

Let

$$T_n = \left\{ \tau : |\tau - \tau_0| \leq C n^{-1} \log \log n \right\}$$

for some constant $C > 0$. Note that we assume that if $\delta_0 \neq 0$, then

$$|\widehat{\tau} - \tau_0| = O_P(n^{-1}),$$

which implies that $\widehat{\tau} \in T_n$ w.p.a.1. Furthermore, note that

$$
\begin{aligned}
R(\alpha_0, \tau) = \, & \mathbb{E}\left(\left[\rho\left(Y, X^T\theta_0\right) - \rho\left(Y, X^T\beta_0\right)\right] \mathbf{1}\{\tau < Q \leq \tau_0\}\right) \\
& + \mathbb{E}\left(\left[\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right] \mathbf{1}\{\tau_0 < Q \leq \tau\}\right).
\end{aligned}
\qquad \text{(E.16)}
$$

Combining the fact that the objective function is Liptschitz continuous by Assumptions 7 (i) with Assumption 1, we have that

$$
\begin{aligned}
\sup_{\tau \in T_n} |R(\alpha_0, \tau)| & \leq L \sup_{\tau \in T_n} \left[\mathbb{E}\left(|X^T\delta_0| \mathbf{1}\{\tau < Q \leq \tau_0\}\right) + \mathbb{E}\left(|X^T\delta_0| \mathbf{1}\{\tau_0 < Q \leq \tau\}\right) \right] \\
& = O\left(|\delta_0|_1 \, n^{-1} \log \log n\right) \\
& = o\left(|\delta_0|_1 \, \omega_n^2\right).
\end{aligned}
$$

Thus, $M^* = O\left(|\alpha_0|_1\right) = O(s)$.

(**Case 2:** $\delta_0 = 0$) Redefine M^* with $T_n = \mathcal{T}$ as the maximum over the whole parameter space for τ. Note that when $\delta_0 = 0$, we have that $R(\alpha_0, \tau) = 0$ and $M^* = O(|\alpha_0|_1) = O(s)$. Therefore, the desired result follows immediately.

E.4 Proof of Theorem 4.1

Remark E.1. We first briefly provide the logic behind the proof of Theorem 4.1 here. Note that for all $\alpha \equiv (\beta^T, \delta^T)^T \in \mathbb{R}^{2p}$ and $\theta \equiv \beta + \delta$, the excess risk has the following decomposition: when $\tau_1 < \tau_0$,

$$
\begin{aligned}
R(\alpha, \tau_1) = {} & \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \leq \tau_1\right\}\right) \\
& + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{Q > \tau_0\right\}\right) \\
& + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{\tau_1 < Q \leq \tau_0\right\}\right),
\end{aligned}
\tag{E.17}
$$

and when $\tau_2 > \tau_0$,

$$
\begin{aligned}
R(\alpha, \tau_2) = {} & \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right] 1\left\{Q \leq \tau_0\right\}\right) \\
& + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{Q > \tau_2\right\}\right) \\
& + \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\left\{\tau_0 < Q \leq \tau_2\right\}\right).
\end{aligned}
\tag{E.18}
$$

The key observations are that all the six terms in the above decompositions are non-negative, and are stochastically negligible when taking $\alpha = \breve{\alpha}$, and $\tau_1 = \breve{\tau}$ if $\breve{\tau} < \tau_0$ or $\tau_2 = \breve{\tau}$ if $\breve{\tau} > \tau_0$. This follows from the risk consistency of $R(\breve{\alpha}, \breve{\tau})$. Then, the identification conditions for α_0 and τ_0 (Assumptions 7 (ii)-(iv)), along with Assumption 6 (i), are useful to show that the risk consistency implies the consistency of $\breve{\tau}$.

Proof of Theorem 4.1. Recall from (E.18) that for all $\alpha = (\beta^T, \delta^T)^T \in \mathbb{R}^{2p}$ and $\theta = \beta + \delta$, the excess risk has the following decomposition: when $\tau > \tau_0$,

$$R\left(\alpha,\tau\right) = \mathbb{E}\left(\left[\rho\left(Y,X^T\beta\right) - \rho\left(Y,X^T\beta_0\right)\right]\mathbb{1}\left\{Q \leq \tau_0\right\}\right)$$
$$+ \mathbb{E}\left(\left[\rho\left(Y,X^T\theta\right) - \rho\left(Y,X^T\theta_0\right)\right]\mathbb{1}\left\{Q > \tau\right\}\right) \qquad \text{(E.19)}$$
$$+ \mathbb{E}\left(\left[\rho\left(Y,X^T\beta\right) - \rho\left(Y,X^T\theta_0\right)\right]\mathbb{1}\left\{\tau_0 < Q \leq \tau\right\}\right).$$

We split the proof into five steps.

Step 1: All the three terms on the right hand side (RHS) of (E.19) are nonnegative. As a consequence, all the three terms on the RHS of (E.19) are bounded by $R(\alpha,\tau)$.

Proof of Step 1. Step 1 is implied by the condition that $\mathbb{E}[\rho(Y,X(\tau_0)^T\alpha) - \rho(Y,X(\tau_0)^T\alpha_0)|Q] \geq 0$ a.s. for all $\alpha \in \mathcal{A}$. To see this, the first two terms are nonnegative by simply multiplying $\mathbb{E}[\rho(Y,X(\tau_0)^T\alpha) - \rho(Y,X(\tau_0)^T\alpha_0)|Q] \geq 0$ with $\mathbb{1}\{Q \leq \tau_0\}$ and $\mathbb{1}\{Q > \tau\}$ respectively. To show that the third term is nonnegative for all $\beta \in \mathbb{R}^p$ and $\tau > \tau_0$, set $\alpha = (\beta/2, \beta/2)$ in the inequality $\mathbb{1}\{\tau_0 < Q \leq \tau\}\mathbb{E}[\rho(Y,X(\tau_0)^T\alpha) - \rho(Y,X(\tau_0)^T\alpha_0)|Q] \geq 0$. Then we have that

$$\mathbb{1}\{\tau_0 < Q \leq \tau\}\mathbb{E}[\rho(Y,X^T(\beta/2 + \beta/2)) - \rho(Y,X^T\theta_0)|Q] \geq 0,$$

which yields the nonnegativeness of the third term. ∎

Step 2: Let $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. Prove:

$$\mathbb{E}\left[|X^T(\beta - \beta_0)|\mathbb{1}\{Q \leq \tau_0\}\right] \leq \frac{1}{\eta^* r^*}R(\alpha,\tau) \vee \left[\frac{1}{\eta^*}R(\alpha,\tau)\right]^{1/2}.$$

Proof of Step 2. Recall that

$$r_1(\eta) \equiv \sup_r \left\{r : \mathbb{E}\left(\left[\rho\left(Y,X^T\beta\right) - \rho\left(Y,X^T\beta_0\right)\right]\mathbb{1}\left\{Q \leq \tau_0\right\}\right)\right.$$
$$\left. \geq \eta\mathbb{E}[(X^T(\beta - \beta_0))^2\mathbb{1}\{Q \leq \tau_0\}] \text{ for all } \beta \in \mathcal{B}(\beta_0, r)\right\}.$$

For notational simplicity, write

$$\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}] \equiv \|\beta - \beta_0\|_q^2,$$

and

$$F(\delta) \equiv \mathbb{E}\left(\left[\rho\left(Y, X^T(\beta_0 + \delta)\right) - \rho\left(Y, X^T \beta_0\right)\right] 1\{Q \le \tau_0\}\right).$$

Note that $F(\beta - \beta_0) = \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T \beta_0\right)\right] 1\{Q \le \tau_0\}\right)$, and $\beta \in \mathcal{B}(\beta_0, r)$ if and only if $\|\beta - \beta_0\|_q \le r$.

For any β, if $\|\beta - \beta_0\|_q \le r_1(\eta^*)$, then by the definition of $r_1(\eta^*)$, we have:

$$F(\beta - \beta_0) \ge \eta^* \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}].$$

If $\|\beta - \beta_0\|_q > r_1(\eta^*)$, let $t = r_1(\eta^*)\|\beta - \beta_0\|_q^{-1} \in (0, 1)$. Since $F(\cdot)$ is convex, and $F(0) = 0$, we have $F(\beta - \beta_0) \ge t^{-1}F(t(\beta - \beta_0))$. Moreover, define

$$\check{\beta} = \beta_0 + r_1(\eta^*)\frac{\beta - \beta_0}{\|\beta - \beta_0\|_q},$$

then $\|\check{\beta} - \beta_0\|_q = r_1(\eta^*)$ and $t(\beta - \beta_0) = \check{\beta} - \beta_0$. Hence still by the definition of $r_1(\eta^*)$,

$$F(\beta - \beta_0) \ge \frac{1}{t}F(\check{\beta} - \beta_0) \ge \frac{\eta^*}{t}\mathbb{E}[(X^T(\check{\beta} - \beta_0))^2 1\{Q \le \tau_0\}] = \eta^* r_1(\eta^*)\|\beta - \beta_0\|_q.$$

Therefore, by Assumption 7 (iii), and Step 1,

$$R(\alpha, \tau) \ge \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T \beta_0\right)\right] 1\{Q \le \tau_0\}\right)$$

$$\ge \eta^* \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}] \wedge \eta^* r^* \{\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}]\}^{1/2}$$

$$\ge \eta^* \left(\mathbb{E}\left[|X^T(\beta - \beta_0)|1\{Q \le \tau_0\}\right]\right)^2 \wedge \eta^* r^* \mathbb{E}\left[|X^T(\beta - \beta_0)|1\{Q \le \tau_0\}\right],$$

where the last inequality follows from Jensen's inequality. ∎

Step 3: For any $r > 0$, w.p.a.1, $\breve{\beta} \in \mathcal{B}(\beta_0, r)$ and $\breve{\theta} \in \mathcal{G}(\theta_0, r)$.

Proof of Step 3. Suppose that $\breve{\tau} > \tau_0$. The proof of Step 2 implies that when $\tau > \tau_0$,

$$\mathbb{E}\left[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}\right] \le \frac{R(\alpha, \tau)^2}{(\eta^* r^*)^2} \vee \frac{R(\alpha, \tau)}{\eta^*}.$$

For any $r > 0$, note that $R(\breve{\alpha}, \breve{\tau}) = o_P(1)$ implies that the event $R(\breve{\alpha}, \breve{\tau}) < r^2$ holds w.p.a.1. Therefore, we have shown that $\breve{\beta} \in \mathcal{B}(\beta_0, r)$.

We now show that $\breve{\theta} \in \mathcal{G}(\theta_0, r)$. When $\tau > \tau_0$, we have that

$$
\begin{aligned}
R(\alpha, \tau) \ge_{(1)} &\ \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{Q > \tau\}\right) \\
= &\ \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{Q > \tau_0\}\right) \\
&- \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{\tau_0 < Q \le \tau\}\right) \\
\ge_{(2)} &\ \eta^* \mathbb{E}\left[|X^T(\theta - \theta_0)|^2 1\{Q > \tau_0\}\right] \wedge \eta^* r^* \left(\mathbb{E}\left[|X^T(\theta - \theta_0)|^2 1\{Q > \tau_0\}\right]\right)^{1/2} \\
&- \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{\tau_0 < Q \le \tau\}\right),
\end{aligned}
$$

where (1) is from (E.18) and (2) can be proved using arguments similar to those used in the proof of Step 2. This implies that

$$\mathbb{E}\left[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}\right] \le \frac{\tilde{R}(\alpha, \tau)^2}{(\eta^* r^*)^2} \vee \frac{\tilde{R}(\alpha, \tau)}{\eta^*},$$

where $\tilde{R}(\alpha, \tau) \equiv R(\alpha, \tau) + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right] 1\{\tau_0 < Q \le \tau\}\right)$. Thus, it suffices to show that $\tilde{R}(\breve{\alpha}, \breve{\tau}) = o_P(1)$ in order to establish that $\breve{\theta} \in \mathcal{G}(\theta_0, r)$. Note that for some

constant $C > 0$,

$$\mathbb{E}\left[(\rho(Y, X^T\theta) - \rho(Y, X^T\theta_0))1\{\tau_0 < Q \leq \tau\}\right]$$

$$\leq_{(1)} L\mathbb{E}\left[|X^T(\theta - \theta_0)|1\{\tau_0 < Q \leq \tau\}\right]$$

$$\leq_{(2)} L|\theta - \theta_0|_1\mathbb{E}\left[\max_{j\leq p}|\tilde{X}_j|1\{\tau_0 < Q \leq \tau\}\right] + L|\theta - \theta_0|_1\mathbb{E}\left[|Q|1\{\tau_0 < Q \leq \tau\}\right]$$

$$\leq_{(3)} L|\theta - \theta_0|_1\mathbb{E}\left[\max_{j\leq p}|\tilde{X}_j|\sup_{\tilde{x}}\mathbb{P}(\tau_0 < Q \leq \tau|\tilde{X} = \tilde{x})\right] + L|\theta - \theta_0|_1\mathbb{E}\left[|Q|1\{\tau_0 < Q \leq \tau\}\right]$$

$$\leq_{(4)} C(\tau - \tau_0)|\theta - \theta_0|_1\mathbb{E}\left\{\left[\max_{j\leq p}|\tilde{X}_j|\right] + 1\right\},$$

where (1) is by the Lipschitz continuity of $\rho(Y, \cdot)$, (2) is from the fact that $|X^T(\theta - \theta_0)| \leq |\theta - \theta_0|_1(\max_{j\leq p}|\tilde{X}_j| + |Q|)$, (3) is by taking the conditional probability, and (4) is from Assumption 5 (ii).

By the expectation-form of the Bernstein inequality (Lemma 14.12 of Bühlmann and van de Geer (2011)), $\mathbb{E}[\max_{j\leq p}|X_j|] \leq K_1 \log(p+1) + \sqrt{2\log(p+1)}$. By (E.27), which will be shown below, $|\breve{\theta} - \theta_0|_1 = O_P(s)$. Hence by (E.23) which will also be shown below, when $\breve{\tau} > \tau_0$,

$$|\breve{\tau} - \tau_0||\breve{\theta} - \theta_0|_1\mathbb{E}[\max_{j\leq p}|X_j|] = O_P(\kappa_n s^2 \log p) = o_P(1).$$

Note that when $\breve{\tau} > \tau_0$, the proofs of (E.23) and (E.27) do not require $\breve{\theta} \in \mathcal{G}(\theta_0, r)$, so there is no problem of applying them here. This implies that $\tilde{R}(\breve{\alpha}, \breve{\tau}) = o_P(1)$.

The same argument yields that w.p.a.1, $\breve{\theta} \in \mathcal{G}(\theta_0, r)$ and $\breve{\beta} \in \mathcal{B}(\beta_0, r)$ when $\breve{\tau} \leq \tau_0$; hence it is omitted to avoid repetition. ∎

Step 4: For any $\epsilon' > 0$ and any $r > 0$, there is an $\varepsilon > 0$ such that for all τ, $\beta \in \mathcal{B}(\beta_0, r)$ and $\theta \in \mathcal{G}(\theta_0, r)$, $R(\alpha, \tau) < \varepsilon$ implies $|\tau - \tau_0| < \epsilon'$.

Proof of Step 4. We first prove that, for any $\epsilon' > 0$, there is $\varepsilon > 0$ such that for all $\tau > \tau_0$, $\beta \in \mathcal{B}(\beta_0, r)$ and $\theta \in \mathcal{G}(\theta_0, r)$, $R(\alpha, \tau) < \varepsilon$ implies that $\tau < \tau_0 + \epsilon'$.

Suppose that $R(\alpha, \tau) < \varepsilon$. Applying the triangle inequality, for all β and $\tau > \tau_0$,

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]$$

$$\leq \left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\right| \qquad \text{(E.20)}$$

$$+ \left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\right|.$$

First, note that the first term on the RHS of (E.20) is the third term on the RHS of (E.19), hence is bounded by $R(\alpha, \tau) < \varepsilon$.

We now consider the second term on the RHS of (E.20). Assumption 6 (i) implies that for all $\beta \in \mathcal{B}(\beta_0, r)$ and $\theta \in \mathcal{G}(\theta_0, r)$,

$$C_2^* \mathbb{E}\left[\left|X^T\beta\right| 1\left\{Q > \tau_0\right\}\right] \leq \mathbb{E}\left[\left|X^T\beta\right| 1\left\{Q \leq \tau_0\right\}\right] \leq C_1^* \mathbb{E}\left[\left|X^T\beta\right| 1\left\{Q > \tau_0\right\}\right]. \qquad \text{(E.21)}$$

It follows from the Lipschitz condition, Step 2, and Assumption 6 (i) that for all $\beta \in \mathcal{B}(\beta_0, r)$,

$$\left|\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\left\{\tau_0 < Q \leq \tau\right\}\right]\right| \leq L\mathbb{E}\left[\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q \leq \tau\right\}\right]$$

$$\leq L\mathbb{E}\left[\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{\tau_0 < Q\right\}\right]$$

$$\leq L\widetilde{C}\,\mathbb{E}\left[\left|X^T\left(\beta - \beta_0\right)\right| 1\left\{Q \leq \tau_0\right\}\right]$$

$$\leq L\widetilde{C}\left\{\varepsilon/(\eta^* r^*) \vee \sqrt{\varepsilon/\eta^*}\right\}$$

$$\equiv C(\varepsilon).$$

Thus, we have shown that (E.20) is bounded by $C(\varepsilon) + \varepsilon$.

For any $\epsilon' > 0$, it follows from Assumptions 7 (ii), 7 (iv) and 5 (iii) (see also Remark

D.2) that there is a $c > 0$ such that if $\tau > \tau_0 + \epsilon'$,

$$c\mathbb{P}\left(\tau_0 < Q \leq \tau_0 + \epsilon'\right) \leq c\mathbb{P}\left(\tau_0 < Q \leq \tau\right)$$
$$\leq \mathbb{E}\left[\left(\rho\left(Y, X^T\beta_0\right) - \rho\left(Y, X^T\theta_0\right)\right)\mathbf{1}\left\{\tau_0 < Q \leq \tau\right\}\right]$$
$$\leq C(\varepsilon) + \varepsilon.$$

Since $\varepsilon \mapsto C(\varepsilon) + \varepsilon$ converges to zero as ε converges to zero, for a given $\epsilon' > 0$ choose a sufficient small $\varepsilon > 0$ such that $C(\varepsilon) + \varepsilon < c\mathbb{P}(\tau_0 < Q \leq \tau_0 + \epsilon')$, so that the above inequality cannot hold. Hence we infer that for this ε, when $R(\alpha, \tau) < \varepsilon$, we must have $\tau < \tau_0 + \epsilon'$.

By the same argument, if $\tau < \tau_0$, then we must have $\tau > \tau_0 - \epsilon'$. Hence, $R(\alpha, \tau) < \varepsilon$ implies $|\tau - \tau_0| < \epsilon'$. ∎

Step 5: $\check{\tau} \xrightarrow{p} \tau_0$.

Proof of Step 5. For the ε chosen in Step 4, consider the event $\{R(\check{\alpha}, \check{\tau}) < \varepsilon\}$, which occurs w.p.a.1, due to Theorem 3.1. On this event, $|\check{\tau} - \tau_0| < \epsilon'$ by Step 4. Because ϵ' is taken arbitrarily, we have proved the consistency of $\check{\tau}$. ∎

∎

E.5 Proof of Theorem 4.2

The proof consists of multiple steps. First, we obtain an intermediate convergence rate for $\check{\tau}$ based on the consistency of the risk and that of $\check{\tau}$. Second, we use the compatibility condition to obtain a tighter bound.

Step 1: Let $\bar{c}_0(\delta_0) \equiv c_0 \inf_{\tau \in \mathcal{T}_0} \mathbb{E}[(X^T\delta_0)^2 | Q = \tau]$, which is bounded away from zero and bounded above due to Assumption 5 (iii). Then $\bar{c}_0(\delta_0)|\check{\tau} - \tau_0| \leq 4R(\check{\alpha}, \check{\tau})$ w.p.a.1. As a result, $|\check{\tau} - \tau_0| = O_P[\kappa_n s / \bar{c}_0(\delta_0)]$.

Proof of Step 1. For any $\tau_0 < \tau$ and $\tau \in \mathcal{T}_0$, and any $\beta \in \mathcal{B}(\beta_0, r)$, $\alpha = (\beta, \delta)$ with arbitrary δ, for some $L, M > 0$ which do not depend on β and τ,

$$\left| \mathbb{E} \left(\rho \left(Y, X^T \beta \right) - \rho \left(Y, X^T \beta_0 \right) \right) 1 \left\{ \tau_0 < Q \leq \tau \right\} \right|$$

$$\leq_{(1)} L\mathbb{E} \left[\left| X^T \left(\beta - \beta_0 \right) \right| 1 \left\{ \tau_0 < Q \leq \tau \right\} \right]$$

$$\leq_{(2)} ML(\tau - \tau_0)\mathbb{E} \left[\left| X^T \left(\beta - \beta_0 \right) \right| 1 \left\{ Q \leq \tau_0 \right\} \right]$$

$$\leq_{(3)} ML(\tau - \tau_0) \left\{ \mathbb{E} \left[\left(X^T \left(\beta - \beta_0 \right) \right)^2 1 \left\{ Q \leq \tau_0 \right\} \right] \right\}^{1/2}$$

$$\leq_{(4)} \left(ML(\tau - \tau_0) \right)^2 / \left(4\eta^* \right) + \eta^* \mathbb{E} \left[\left(X^T \left(\beta - \beta_0 \right) \right)^2 1 \left\{ Q \leq \tau_0 \right\} \right]$$

$$\leq_{(5)} \left(ML(\tau - \tau_0) \right)^2 / \left(4\eta^* \right) + \mathbb{E} \left[\left(\rho \left(Y, X^T \beta \right) - \rho \left(Y, X^T \beta_0 \right) \right) 1 \left\{ Q \leq \tau_0 \right\} \right]$$

$$\leq_{(6)} \left(ML(\tau - \tau_0) \right)^2 / \left(4\eta^* \right) + R(\alpha, \tau),$$

where (1) follows from the Lipschitz condition on the objective function, (2) is by Assumption 6 (ii), (3) is by Jensen's inequality, (4) follows from the fact that $uv \leq v^2 / (4c) + cu^2$ for any $c > 0$, (5) is from Assumption 7 (iii), and (6) is from Step 1 in the proof of Theorem 4.1.

In addition,

$$\left| \mathbb{E} \left[\left(\rho \left(Y, X^T \beta \right) - \rho \left(Y, X^T \beta_0 \right) \right) 1 \left\{ \tau_0 < Q \leq \tau \right\} \right] \right|$$

$$\geq_{(1)} \mathbb{E} \left[\left(\rho \left(Y, X^T \beta_0 \right) - \rho \left(Y, X^T \theta_0 \right) \right) 1 \left\{ \tau_0 < Q \leq \tau \right\} \right]$$

$$- \left| \mathbb{E} \left[\left(\rho \left(Y, X^T \beta \right) - \rho \left(Y, X^T \theta_0 \right) \right) 1 \left\{ \tau_0 < Q \leq \tau \right\} \right] \right|$$

$$\geq_{(2)} \mathbb{E} \left[\left(\rho \left(Y, X^T \beta_0 \right) - \rho \left(Y, X^T \theta_0 \right) \right) 1 \left\{ \tau_0 < Q \leq \tau \right\} \right] - R(\alpha, \tau)$$

$$\geq_{(3)} c_0 \left\{ \inf_{\tau \in \mathcal{T}_0} \mathbb{E}[(X^T \delta_0)^2 | Q = \tau] \right\} (\tau - \tau_0) - R(\alpha, \tau),$$

where (1) is by the triangular inequality, (2) is from (E.18), and (3) is by Assumption 7 (iv). Therefore, we have established that there exists a constant $\tilde{C} > 0$, independent of (α, τ),

such that

$$\bar{c}_0(\delta_0)(\tau - \tau_0) \leq \tilde{C}(\tau - \tau_0)^2 + 2R(\alpha, \tau). \tag{E.22}$$

Note that when $0 < (\tau - \tau_0) < \bar{c}_0(\delta_0)(2\tilde{C})^{-1}$, (E.22) implies that

$$\bar{c}_0(\delta_0)(\tau - \tau_0) \leq \frac{\bar{c}_0(\delta_0)}{2}(\tau - \tau_0) + 2R(\alpha, \tau),$$

which in turn implies that $\tau - \tau_0 \leq \frac{4}{\bar{c}_0(\delta_0)} R(\alpha, \tau)$. By the same argument, when $-\bar{c}_0(\delta_0)(2\tilde{C})^{-1} < (\tau - \tau_0) \leq 0$, we have $\tau_0 - \tau \leq \frac{4}{\bar{c}_0(\delta_0)} R(\alpha, \tau)$ for $\alpha = (\beta, \delta)$, with any $\theta \in \mathcal{G}(\theta_0, r)$ and arbitrary β.

Hence when $\breve{\tau} > \tau_0$, on the event $\breve{\beta} \in \mathcal{B}(\beta_0, r)$, and $\breve{\tau} - \tau_0 < \bar{c}_0(\delta_0)(2\tilde{C})^{-1}$, we have

$$\breve{\tau} - \tau_0 \leq \frac{4}{\bar{c}_0(\delta_0)} R(\breve{\alpha}, \breve{\tau}). \tag{E.23}$$

When $\breve{\tau} \leq \tau_0$, on the event $\breve{\theta} \in \mathcal{G}(\theta_0, r)$, and $\tau_0 - \breve{\tau} < \bar{c}_0(\delta_0)(2\tilde{C})^{-1}$, we have $\tau_0 - \breve{\tau} \leq \frac{4}{\bar{c}_0(\delta_0)} R(\breve{\alpha}, \breve{\tau})$. Hence due to Step 3 in the proof of Theorem 4.1 and the consistency of $\breve{\tau}$, we have

$$|\breve{\tau} - \tau_0| \leq \frac{4}{\bar{c}_0(\delta_0)} R(\breve{\alpha}, \breve{\tau}) \quad \text{w.p.a.1.} \tag{E.24}$$

This also implies $|\breve{\tau} - \tau_0| = O_P[\kappa_n s / \bar{c}_0(\delta_0)]$ in view of the proof of Theorem 3.1. ∎

Step 2: Define $\nu_{1n}(\tau) \equiv \nu_n(\alpha_0, \tau) - \nu_n(\alpha_0, \tau_0)$ and $c_\alpha \equiv \kappa_n\left(|D_0\alpha_0|_1 - \left|\breve{D}\alpha_0\right|_1\right) + |\nu_{1n}(\breve{\tau})|$. Then,

$$R(\breve{\alpha}, \breve{\tau}) + \frac{1}{2}\kappa_n\left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 \leq c_\alpha + 2\kappa_n\left|\breve{D}(\breve{\alpha} - \alpha_0)_J\right|_1 \quad \text{w.p.a.1.} \tag{E.25}$$

Proof of Step 2. Recall the following basic inequality in (E.7):

$$R(\breve{\alpha}, \breve{\tau}) \leq [\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})] - \nu_{1n}(\breve{\tau}) + \kappa_n\left(|D_0\alpha_0|_1 - |\breve{D}\breve{\alpha}|_1\right). \tag{E.26}$$

Now applying Lemma E.1 to $[\nu_n(\alpha_0, \check{\tau}) - \nu_n(\check{\alpha}, \check{\tau})]$ with a_n and b_n replaced by $a_n/2$ and $b_n/2$, we can rewrite the basic inequality in (E.26) by

$$\kappa_n |D_0 \alpha_0|_1 \geq R(\check{\alpha}, \check{\tau}) + \kappa_n \left| \check{D}\check{\alpha} \right|_1 - \frac{1}{2}\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0) \right|_1 - |\nu_{1n}(\check{\tau})| \quad \text{w.p.a.1.}$$

Now adding $\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0) \right|_1$ on both sides of the inequality above and using the fact that $|\alpha_{0j}|_1 - |\check{\alpha}_j|_1 + |(\check{\alpha}_j - \alpha_{0j})|_1 = 0$ for $j \notin J$, we have that

$$\kappa_n \left(|D_0 \alpha_0|_1 - \left| \check{D}\alpha_0 \right|_1 \right) + |\nu_{1n}(\check{\tau})| + 2\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1$$
$$\geq R(\check{\alpha}, \check{\tau}) + \frac{1}{2}\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0) \right|_1 \quad \text{w.p.a.1.}$$

Therefore, we have proved Step 2. ∎

We prove the remaining part of the steps by considering two cases: (i) $\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1 \leq c_\alpha$; (ii) $\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1 > c_\alpha$. We first consider Case (ii).

Step 3: Suppose that $\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1 > c_\alpha$. Then

$$|\check{\tau} - \tau_0| = O_P\left[\kappa_n^2 s / \bar{c}_0(\delta_0)\right] \quad \text{and} \quad |\check{\alpha} - \alpha_0| = O_P(\kappa_n s).$$

Proof of Step 3. By $\kappa_n \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1 > c_\alpha$ and the basic inequality (E.25) in Step 2,

$$6 \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1 \geq \left| \check{D}(\check{\alpha} - \alpha_0) \right|_1 = \left| \check{D}(\check{\alpha} - \alpha_0)_J \right|_1 + \left| \check{D}(\check{\alpha} - \alpha_0)_{J^c} \right|_1, \tag{E.27}$$

which enables us to apply the compatibility condition in Assumption 3.

Recall that $\|Z\|_2 = (EZ^2)^{1/2}$ for a random variable Z. Note that for $s = |J(\alpha_0)|_0$,

$$
\begin{aligned}
& R\left(\breve{\alpha}, \breve{\tau}\right) + \frac{1}{2}\kappa_n \left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)\right|_1 \\
& \leq_{(1)} 3\kappa_n \left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)_J\right|_1 \\
& \leq_{(2)} 3\kappa_n \bar{D} \left\|X(\breve{\tau})^T(\breve{\alpha} - \alpha_0)\right\|_2 \sqrt{s}/\phi \\
& \leq_{(3)} \frac{9\kappa_n^2 \bar{D}^2 s}{2\tilde{c}\phi^2} + \frac{\tilde{c}}{2} \left\|X(\breve{\tau})^T(\breve{\alpha} - \alpha_0)\right\|_2^2,
\end{aligned}
\tag{E.28}
$$

where (1) is from the basic inequality (E.25) in Step 2, (2) is by the compatibility condition (Assumption 3), and (3) is from the inequality that $uv \leq v^2/(2\tilde{c}) + \tilde{c}u^2/2$ for any $\tilde{c} > 0$.

We will show below in Step 4 that there is a constant $C_0 > 0$ such that

$$
\left\|X(\breve{\tau})^T(\breve{\alpha} - \alpha_0)\right\|_2^2 \leq C_0 R(\breve{\alpha}, \breve{\tau}) + C_0 \bar{c}_0(\delta_0)|\breve{\tau} - \tau_0|, \text{ w.p.a.1.} \tag{E.29}
$$

Recall that by (E.24), $\bar{c}_0(\delta_0)\left|\breve{\tau} - \tau_0\right| \leq 4R\left(\breve{\alpha}, \breve{\tau}\right)$. Hence, (E.28) with $\tilde{c} = (5C_0)^{-1}$ implies that

$$
R\left(\breve{\alpha}, \breve{\tau}\right) + \kappa_n \left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)\right|_1 \leq \frac{9\kappa_n^2 \bar{D}^2 s}{\tilde{c}\phi^2}. \tag{E.30}
$$

By (E.30) and (E.24), $|\breve{\tau} - \tau_0| = O_P\left[\kappa_n^2 s/\bar{c}_0(\delta_0)\right]$. Also, by (E.30), $|\breve{\alpha} - \alpha_0| = O_P\left(\kappa_n s\right)$ since $D(\breve{\tau}) \geq \underline{D}$ w.p.a.1 by Assumption 1 (iv). ∎

Step 4: There is a constant $C_0 > 0$ such that $\left\|X(\breve{\tau})^T(\breve{\alpha} - \alpha_0)\right\|_2^2 \leq C_0 R(\breve{\alpha}, \breve{\tau}) + C_0 \bar{c}_0(\delta_0)|\breve{\tau} - \tau_0|$, w.p.a.1.

Proof Step 4. Note that

$$
\begin{aligned}
\left\|X(\tau)^T(\alpha - \alpha_0)\right\|_2^2 \leq{} & 2\left\|X(\tau)^T\alpha - X(\tau_0)^T\alpha\right\|_2^2 \\
& + 4\left\|X(\tau_0)^T\alpha - X(\tau_0)^T\alpha_0\right\|_2^2 + 4\left\|X(\tau_0)^T\alpha_0 - X(\tau)^T\alpha_0\right\|_2^2.
\end{aligned}
\tag{E.31}
$$

We bound the three terms on the right hand side of (E.31). When $\tau > \tau_0$, there is a constant

$C_1 > 0$ such that

$$\left\|X(\tau)^T\alpha - X(\tau_0)^T\alpha\right\|_2^2$$

$$= \mathbb{E}\left[(X^T\delta)^2 1\{\tau_0 \leq Q < \tau\}\right]$$

$$= \int_{\tau_0}^{\tau} \mathbb{E}\left[(X^T\delta)^2\big|Q = t\right] dF_Q(t)$$

$$\leq 2 \int_{\tau_0}^{\tau} \mathbb{E}\left[(X^T\delta_0)^2\big|Q = t\right] dF_Q(t) + 2 \int_{\tau_0}^{\tau} \mathbb{E}\left[(X^T(\delta - \delta_0))^2\big|Q = t\right] dF_Q(t)$$

$$\leq C_1\bar{c}_0(\delta_0)(\tau - \tau_0),$$

where the last inequality is by Assumptions 1, 5 (ii), 5 (iii), and 6 (ii).

Similarly, $\left\|X(\tau_0)^T\alpha_0 - X(\tau)^T\alpha_0\right\|_2^2 = \mathbb{E}\left[(X^T\delta_0)^2 1\{\tau_0 \leq Q < \tau\}\right] \leq C_1\bar{c}_0(\delta_0)(\tau - \tau_0)$.

Hence, the first and third terms of the right hand side of of (E.31) are bounded by $6C_1\bar{c}_0(\delta_0)(\tau - \tau_0)$.

To bound the second term, note that there exists a constant $C_2 > 0$ such that

$$\left\|X(\tau_0)^T\alpha - X(\tau_0)^T\alpha_0\right\|_2^2$$

$$=_{(1)} \mathbb{E}\left[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}\right] + \mathbb{E}\left[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}\right]$$

$$\leq_{(2)} (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{Q > \tau_0\}\right]$$

$$+ (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \leq \tau_0\}\right]$$

$$\leq_{(3)} (\eta^*)^{-1}R(\alpha, \tau) + (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{\tau_0 < Q \leq \tau\}\right]$$

$$\leq_{(4)} (\eta^*)^{-1}R(\alpha, \tau) + (\eta^*)^{-1}L\mathbb{E}\left[|X^T(\theta - \theta_0)|1\{\tau_0 < Q \leq \tau\}\right]$$

$$=_{(5)} (\eta^*)^{-1}R(\alpha, \tau) + (\eta^*)^{-1}L \int_{\tau_0}^{\tau} \mathbb{E}\left[|X^T(\theta - \theta_0)|\big|Q = t\right] dF_Q(t)$$

$$\leq_{(6)} (\eta^*)^{-1}R(\alpha, \tau) + C_3(\tau - \tau_0),$$

where (1) is simply an identity, (2) from Assumption 7 (iii), (3) is due to (E.19): namely,

$$\mathbb{E}\left[\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\theta_0\right)\right) 1\{Q > \tau\}\right] + \mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \leq \tau_0\}\right] \leq R(\alpha, \tau),$$

(4) is by the Lipschitz continuity of $\rho(Y, \cdot)$, (5) is by rewriting the expectation term, and (6) is by Assumptions 1 (ii) and 6 (ii). Therefore, we have shown that $\left\| X(\tau)^T (\alpha - \alpha_0) \right\|_2^2 \leq C_0 R(\alpha, \tau) + C_0 \bar{c}_0(\delta_0)(\tau - \tau_0)$ for some constant $C_0 > 0$. The case of $\tau \leq \tau_0$ can be proved using the same argument. Hence, setting $\tau = \check{\tau}$, and $\alpha = \check{\alpha}$, we obtain the desired result. ∎

Step 5: We now consider Case (i). Suppose that $\kappa_n \left| \check{D} (\check{\alpha} - \alpha_0)_J \right|_1 \leq c_\alpha$. Then

$$|\check{\tau} - \tau_0| = O_P \left[\kappa_n^2 s / \bar{c}_0(\delta_0) \right] \quad \text{and} \quad |\check{\alpha} - \alpha_0| = O_P \left(\kappa_n s \right).$$

Proof of Step 5. Recall that X_{ij} is the jth element of X_i, where $i \leq n, j \leq p$. By Assumption 1 and Step 1,

$$\sup_{1 \leq j \leq p} \frac{1}{n} \sum_{i=1}^n |X_{ij}|^2 \left| 1 \left(Q_i < \check{\tau} \right) - 1 \left(Q_i < \tau_0 \right) \right| = O_P \left[\kappa_n s / \bar{c}_0(\delta_0) \right].$$

By the mean value theorem,

$$
\begin{aligned}
& \kappa_n \left| |D_0 \alpha_0|_1 - \left| \check{D} \alpha_0 \right|_1 \right| \\
& \leq \kappa_n \sum_{j=1}^p \left(\frac{4}{n} \sum_{i=1}^n |X_{ij} 1 \{ Q_i > \bar{\tau} \}|^2 \right)^{-1/2} \left| \delta_0^{(j)} \right| \frac{1}{n} \sum_{i=1}^n |X_{ij}|^2 \left| 1 \{ Q_i > \check{\tau} \} - 1 \{ Q_i > \tau_0 \} \right| \\
& = O_P \left[\kappa_n^2 s |J(\delta_0)|_0 / \bar{c}_0(\delta_0) \right].
\end{aligned}
\tag{E.32}
$$

Here, recall that $\bar{\tau}$ is the right-end point of $\mathcal{T}$ and $|J(\delta_0)|_0$ is the dimension of nonzero elements of δ_0.

Due to Step 1 and (E.3) in Lemma E.1,

$$|\nu_{1n} (\check{\tau})| = O_P \left[\frac{|\delta_0|_2}{\sqrt{\bar{c}_0(\delta_0)}} (\kappa_n s / n)^{1/2} \right]. \tag{E.33}$$

Thus, under Case (i), we have that, by (E.24), (E.25), (E.32), and (E.33),

$$
\begin{aligned}
\frac{\bar{c}_0(\delta_0)}{4} \left| \check{\tau} - \tau_0 \right| &\leq \frac{\kappa_n}{2} \left| \check{D} \left(\check{\alpha} - \alpha_0 \right) \right|_1 + R \left(\check{\alpha}, \check{\tau} \right) \\
&\leq 3\kappa_n \left(\left| D_0 \alpha_0 \right|_1 - \left| \check{D} \alpha_0 \right|_1 \right) + 3 \left| \nu_{1n} \left(\check{\tau} \right) \right| \\
&= O_P \left(\kappa_n^2 s^2 \right) + O_P \left[s^{1/2} \left(\kappa_n s / n \right)^{1/2} \right],
\end{aligned}
\tag{E.34}
$$

where the last equality uses the fact that $|J(\delta_0)|_0 / \bar{c}_0(\delta_0) = O(s)$ and $|\delta_0|_2 / \sqrt{\bar{c}_0(\delta_0)} = O(s^{1/2})$ at most (both could be bounded in some cases).

Therefore, we now have an improved rate of convergence in probability for $\check{\tau}$ from $r_{n0,\tau} \equiv \kappa_n s$ to $r_{n1,\tau} \equiv [\kappa_n^2 s^2 + s^{1/2}(\kappa_n s / n)^{1/2}]$. Repeating the arguments identical to those to prove (E.32) and (E.33) yields that

$$
\kappa_n \left| \left| D_0 \alpha_0 \right|_1 - \left| \check{D} \alpha_0 \right|_1 \right| = O_P \left[r_{n1,\tau} \kappa_n s \right] \quad \text{and} \quad \left| \nu_{1n} \left(\check{\tau} \right) \right| = O_P \left[s^{1/2} \left(r_{n1,\tau} / n \right)^{1/2} \right].
$$

Plugging these improved rates into (E.34) gives

$$
\begin{aligned}
\bar{c}_0(\delta_0) \left| \check{\tau} - \tau_0 \right| &= O_P \left(\kappa_n^3 s^3 \right) + O_P \left[s^{1/2} (\kappa_n s)^{3/2} / n^{1/2} \right] + O_P \left(\kappa_n s^{3/2} / n^{1/2} \right) + O_P \left[s^{3/4} (\kappa_n s)^{1/4} / n^{3/4} \right] \\
&= O_P \left(\kappa_n^2 s^{3/2} \right) + O_P \left[s^{3/4} (\kappa_n s)^{1/4} / n^{3/4} \right] \\
&\equiv O_P (r_{n2,\tau}),
\end{aligned}
$$

where the second equality comes from the fact that the first three terms are $O_P \left(\kappa_n^2 s^{3/2} \right)$ since $\kappa_n s^{3/2} = o(1)$, $\kappa_n n / s \to \infty$, and $\kappa_n \sqrt{n} \to \infty$ in view of the assumption that $\kappa_n s^2 \log p = o(1)$. Repeating the same arguments again with the further improved rate $r_{n2,\tau}$, we have that

$$
\left| \check{\tau} - \tau_0 \right| = O_P \left(\kappa_n^2 s^{5/4} \right) + O_P \left[s^{7/8} (\kappa_n s)^{1/8} / n^{7/8} \right] \equiv O_P (r_{n3,\tau}).
$$

Thus, repeating the same arguments k times yields

$$\bar{c}_0(\delta_0)\left|\breve{\tau} - \tau_0\right| = O_P\left(\kappa_n^2 s^{1+2^{-k}}\right) + O_P\left[s^{(2^k-1)/2^k}(\kappa_n s)^{1/2^k}/n^{(2^k-1)/2^k}\right] \equiv O_P(r_{nk,\tau}).$$

Then letting $k \to \infty$ gives the desired result that $\bar{c}_0(\delta_0)\left|\breve{\tau} - \tau_0\right| = O_P\left(\kappa_n^2 s\right)$. Finally, the same iteration based on (E.34) gives $\left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)\right| = o_P\left(\kappa_n s\right)$, which proves the desired result since $D(\breve{\tau}) \geq \underline{D}$ w.p.a.1 by Assumption 1 (iv). $\blacksquare$

E.6 Proof of Theorem 4.3

Proof of Theorem 4.3. The asymptotic property of $\widetilde{\tau}$ is well-known in the literature (see Lemma E.3 below for its asymptotic distribution). Specifically, we can apply Theorem 3.4.1 of van der Vaart and Wellner (1996) (by defining the criterion $\mathbb{M}_n\left(\cdot\right) \equiv R_n^*\left(\cdot\right)$, $M_n\left(\cdot\right) \equiv \mathbb{E}R_n^*\left(\cdot\right) = R(\alpha_0, \tau)$, the distance function $d\left(\tau, \tau_0\right) \equiv \left|\tau - \tau_0\right|^{1/2}$, and $\phi_n\left(\delta\right) \equiv \delta$) to characterize the convergence rate of $\widetilde{\tau}$, which results in the super-consistency in the sense that $\widetilde{\tau} - \tau_0 = O_P(n^{-1})$. See e.g. Section 14.5 of Kosorok (2008).

Furthermore, it is worth noting that the same theorem also implies that if

$$[R_n^*\left(\widehat{\tau}\right) - R_n^*\left(\tau_0\right)] - [R_n\left(\breve{\alpha}, \widehat{\tau}\right) - R_n\left(\breve{\alpha}, \tau_0\right)] = O_P(r_n^{-2}) \tag{E.35}$$

for some sequence r_n satisfying $r_n^2\phi_n\left(r_n^{-1}\right) \leq \sqrt{n}$, then

$$r_n d\left(\widehat{\tau}, \tau_0\right) = O_P\left(1\right).$$

This is because

$$R_n^*(\widehat{\tau}) = R_n^*(\widehat{\tau}) - [R_n(\breve{\alpha}, \widehat{\tau}) - R_n(\breve{\alpha}, \tau_0) + R_n^*(\tau_0)] + [R_n(\breve{\alpha}, \widehat{\tau}) - R_n(\breve{\alpha}, \tau_0) + R_n^*(\tau_0)]$$

$$\leq_{(1)} R_n^*(\widehat{\tau}) - [R_n(\breve{\alpha}, \widehat{\tau}) - R_n(\breve{\alpha}, \tau_0) + R_n^*(\tau_0)] + [R_n(\breve{\alpha}, \tau_0) - R_n(\breve{\alpha}, \tau_0) + R_n^*(\tau_0)]$$

$$=_{(2)} \{[R_n^*(\widehat{\tau}) - R_n^*(\tau_0)] - [R_n(\breve{\alpha}, \widehat{\tau}) - R_n(\breve{\alpha}, \tau_0)]\} + R_n^*(\tau_0)$$

$$=_{(3)} O_P(r_n^{-2}) + R_n^*(\tau_0),$$

where inequality (1) uses the fact that $\widehat{\tau}$ is a minimizer of $R_n(\breve{\alpha}, \tau)$, equality (2) follows since $R_n(\breve{\alpha}, \tau_0) - R_n(\breve{\alpha}, \tau_0) + R_n^*(\tau_0) = R_n^*(\tau_0)$, and equality (3) comes from (E.35).

Then, note that we can set $r_n^{-2} = a_n s_n \log(np)$ with $s_n = 1$ and $a_n = \kappa_n s \log n$ due to Lemma E.2 and the rate of convergence $\breve{\alpha} - \alpha_0 = O_P(\kappa_n s)$ given by Theorem 4.2. Next, we will apply a chaining argument to obtain the convergence rate of $\widehat{\tau}$ by repeatedly verifying the condition $R_n^*(\widehat{\tau}) \leq R_n^*(\tau_0) + O_P(r_n^{-2})$, with an iteratively improved rate r_n. Applying Theorem 3.4.1 of van der Vaart and Wellner (1996) with $r_n = (a_n \log(np))^{-1/2}$, we have

$$\widehat{\tau} - \tau_0 = O_P(a_n \log(np)) = O_P(\kappa_n s \log n \log(np)).$$

Next, we reset $s_n = \kappa_n s (\log n)^2 \log(np)$ and $a_n = \kappa_n s \log n$ to apply Lemma E.2 again and then Theorem 3.4.1 of van der Vaart and Wellner (1996) with $r_n = (s_n a_n \log(np))^{-1/2}$. It follows that

$$\widehat{\tau} - \tau_0 = O_P\left([\kappa_n s]^2 (\log n)^3 (\log(np))^2\right).$$

In the next step, we set $r_n = \sqrt{n}$ since it should satisfy the constraint that $r_n^2 \phi_n(r_n^{-1}) \leq \sqrt{n}$ as well. Then, we conclude that $\widehat{\tau} = \tau_0 + O_P(n^{-1})$. Furthermore, in view of Lemma E.2, $\widehat{\tau} = \tau_0 + O_P(n^{-1})$ implies that the asymptotic distribution of $n(\widehat{\tau} - \tau_0)$ is identical to $n(\widetilde{\tau} - \tau_0)$ since each of them is characterized by the minimizer of the weak limit of $n(R_n(\alpha, \tau_0 + tn^{-1}) - R_n(\alpha, \tau_0))$ with $\alpha = \breve{\alpha}$ and $\alpha = \alpha_0$, respectively. That is, the weak limits of the processes are identical due to Lemma E.2. Therefore, we have proved the first

81

conclusion of the theorem. Lemma E.3 establishes the second conclusion. ∎

Lemma E.2. *Suppose that* $\alpha \in \mathcal{A}_n \equiv \left\{ \alpha = \left(\beta^T, \delta^T \right)^T : |\alpha - \alpha_0|_1 \leq K a_n \right\}$ *and* $\tau \in \mathcal{T}_n \equiv$ $\{|\tau - \tau_0| \leq K s_n\}$ *for some* $K < \infty$ *and for some sequences* a_n *and* s_n *as* $n \to \infty$. *Then,*

$$\sup_{\alpha \in \mathcal{A}_n, \tau \in \mathcal{T}_n} \left| \{R_n(\alpha, \tau) - R_n(\alpha, \tau_0)\} - \{R_n(\alpha_0, \tau) - R_n(\alpha_0, \tau_0)\} \right| = O_P[a_n s_n \log(np)].$$

Proof of Lemma E.2. Noting that

$$\rho \left(Y_i, X_i^T \beta + X_i^T \delta 1\{Q_i > \tau\}\right) = \rho \left(Y_i, X_i^T \beta\right) 1\{Q_i \leq \tau\} + \rho \left(Y_i, X_i^T \beta + X_i^T \delta\right) 1\{Q_i > \tau\},$$

we have, for $\tau > \tau_0$,

$$\begin{aligned}
D_n(\alpha, \tau) &:= \{R_n(\alpha, \tau) - R_n(\alpha, \tau_0)\} - \{R_n(\alpha_0, \tau) - R_n(\alpha_0, \tau_0)\} \\
&= \frac{1}{n} \sum_{i=1}^n \left[\rho \left(Y_i, X_i^T \beta\right) - \rho \left(Y_i, X_i^T \beta_0\right) \right] 1\{\tau_0 < Q_i \leq \tau\} \\
&\quad - \frac{1}{n} \sum_{i=1}^n \left[\rho \left(Y_i, X_i^T \theta\right) - \rho \left(Y_i, X_i^T \theta_0\right) \right] 1\{\tau_0 < Q_i \leq \tau\} \\
&=: D_{n1}(\alpha, \tau) - D_{n2}(\alpha, \tau).
\end{aligned}$$

However, the Lipschitz property of ρ yields that

$$\begin{aligned}
|D_{n1}(\alpha, \tau)| &= \left| \frac{1}{n} \sum_{i=1}^n \left[\rho \left(Y_i, X_i^T \beta\right) - \rho \left(Y_i, X_i^T \beta_0\right) \right] 1\{\tau_0 < Q_i \leq \tau\} \right| \\
&\leq L \max_{i,j} |X_{ij}| \, |\beta - \beta_0|_1 \frac{1}{n} \sum_{i=1}^n 1\{\tau_0 < Q_i \leq \tau\} \\
&= O_P[\log(np) \cdot a_n \cdot s_n] \quad \text{uniformly in } (\alpha, \tau) \in \mathcal{A}_n \times \mathcal{T}_n,
\end{aligned}$$

where $\log(np)$ term comes from the Bernstein inequality and the s_n term follows from the fact that $\mathbb{E} \left| \frac{1}{n} \sum_{i=1}^n 1\{\tau_0 < Q_i \leq \tau\} \right| = \mathbb{E} 1\{\tau_0 < Q_i \leq \tau\} \leq C \cdot K s_n$ due to the boundedness of the density of Q_i around τ_0. The other term $D_{n2}(\alpha, \tau)$ can be bounded similarly. The

case of $\tau < \tau_0$ can be treated analogously and hence details are omitted. ∎

Lemma E.3. *We have that $n(\widetilde{\tau} - \tau_0)$ converges in distribution to the smallest minimizer of a compound Poisson process defined in Theorem 4.3.*

Proof of Lemma E.3. The convergence rate of $\widetilde{\tau}$ is standard as commented in the beginning of the proof of Theorem 4.3 and thus details are omitted here. We present the characterization of the asymptotic distribution for the given convergence rate n.

Recall that $\rho(t, s) = \dot{\rho}(t - s)$, where $\dot{\rho}(t) = t(\gamma - 1\{t \leq 0\})$. Note that

$$
\begin{aligned}
&nR_n^*(\tau) \\
&= \sum_{i=1}^n \dot{\rho}\left(Y_i - X_i^T\beta_0 - X_i^T\delta_0 1\{Q_i > \tau\}\right) - \dot{\rho}\left(Y_i - X_i^T\beta_0 - X_i^T\delta_0 1\{Q_i > \tau_0\}\right) \\
&= \sum_{i=1}^n \left[\dot{\rho}\left(U_i - X_i^T\delta_0\left(1\{Q_i > \tau\} - 1\{Q_i > \tau_0\}\right)\right) - \dot{\rho}(U_i)\right]\left(1\{\tau < Q_i \leq \tau_0\} + 1\{\tau_0 < Q_i \leq \tau\}\right) \\
&= \sum_{i=1}^n \left[\dot{\rho}\left(U_i - X_i^T\delta_0\right) - \dot{\rho}(U_i)\right] 1\{\tau < Q_i \leq \tau_0\} \\
&\quad + \sum_{i=1}^n \left[\dot{\rho}\left(U_i + X_i^T\delta_0\right) - \dot{\rho}(U_i)\right] 1\{\tau_0 < Q_i \leq \tau\}.
\end{aligned}
$$

Thus, the asymptotic distribution of $n(\widetilde{\tau} - \tau_0)$ is characterized by the smallest minimizer of the weak limit of

$$
M_n(h) = \sum_{i=1}^n \dot{\rho}_{1i} 1\left\{\tau_0 + \frac{h}{n} < Q_i \leq \tau_0\right\} + \sum_{i=1}^n \dot{\rho}_{2i} 1\left\{\tau_0 < Q_i \leq \tau_0 + \frac{h}{n}\right\}
$$

for $|h| \leq K$ for some large K, where $\dot{\rho}_{1i} \equiv \dot{\rho}\left(U_i - X_i^T\delta_0\right) - \dot{\rho}(U_i)$ and $\dot{\rho}_{2i} \equiv \dot{\rho}\left(U_i + X_i^T\delta_0\right) - \dot{\rho}(U_i)$. The weak limit of the empirical process $M_n(\cdot)$ is well developed in the literature, (see e.g. Pons (2003); Kosorok and Song (2007); Lee and Seo (2008)) and the argmax continuous mapping theorem by Seijo and Sen (2011b) yields the asymptotic distribution, namely the smallest minimizer of a compound Poisson process, which is defined in Theorem 4.3. ∎

E.7 Proof of Theorem 4.4

Let $\widehat{D} \equiv D\left(\widehat{\tau}\right)$. It follows from the definition of $\widehat{\alpha}$ in (2.5) that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\widehat{\tau})^T\widehat{\alpha}) + \omega_n|\widehat{D}\widehat{\alpha}|_1 \leq \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i, X_i(\widehat{\tau})^T\alpha_0) + \omega_n|\widehat{D}\alpha_0|_1.$$

From this, we obtain the following inequality

$$R(\widehat{\alpha}, \widehat{\tau}) \leq [\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})] + R(\alpha_0, \widehat{\tau}) + \omega_n|\widehat{D}\alpha_0|_1 - \omega_n|\widehat{D}\widehat{\alpha}|_1. \qquad \text{(E.36)}$$

Now applying Lemma E.1 to $[\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widehat{\alpha}, \widehat{\tau})]$, we rewrite the basic inequality in (E.36) by

$$\omega_n\left|\widehat{D}\alpha_0\right|_1 \geq R(\widehat{\alpha}, \widehat{\tau}) + \omega_n\left|\widehat{D}\widehat{\alpha}\right|_1 - \frac{1}{2}\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 - |R(\alpha_0, \widehat{\tau})| \quad \text{w.p.a.1.}$$

As before, adding $\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1$ on both sides of the inequality above and using the fact that $|\alpha_{0j}|_1 - |\widehat{\alpha}_j|_1 + |(\widehat{\alpha}_j - \alpha_{0j})|_1 = 0$ for $j \notin J$, we have that

$$R\left(\widehat{\alpha}, \widehat{\tau}\right) + \frac{1}{2}\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)\right|_1 \leq |R(\alpha_0, \widehat{\tau})| + 2\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)_J\right|_1 \quad \text{w.p.a.1.} \qquad \text{(E.37)}$$

As in the proof of Theorem 4.2, we consider two cases: (i) $\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)_J\right|_1 \leq |R(\alpha_0, \widehat{\tau})|$; (ii) $\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)_J\right|_1 > |R(\alpha_0, \widehat{\tau})|$. We first consider case (ii). Recall that $\|Z\|_2 = (EZ^2)^{1/2}$ for a random variable Z. It follows from the compatibility condition (Assumption 3) and the same arguments as in (E.28) that

$$\begin{aligned}
\omega_n\left|\widehat{D}\left(\widehat{\alpha} - \alpha_0\right)_J\right|_1 &\leq \omega_n\bar{D}\left\|X(\widehat{\tau})^T(\widehat{\alpha} - \alpha_0)\right\|_2 \sqrt{s}/\phi \\
&\leq \frac{\omega_n^2\bar{D}^2 s}{2\tilde{c}\phi^2} + \frac{\tilde{c}}{2}\left\|X(\widehat{\tau})^T(\widehat{\alpha} - \alpha_0)\right\|_2^2
\end{aligned} \qquad \text{(E.38)}$$

for any $\tilde{c} > 0$. Recall that $\bar{c}_0(\delta_0) \equiv c_0\inf_{\tau \in \mathcal{T}_0}\mathbb{E}[(X^T\delta_0)^2|Q = \tau]$. As in Step 5 of the proof of

Theorem 4.2, there is a constant $C_0 > 0$ such that

$$\left\| X(\widehat{\tau})^T(\widehat{\alpha} - \alpha_0) \right\|_2^2 \leq C_0 R(\widehat{\alpha}, \widehat{\tau}) + C_0 \bar{c}_0(\delta_0)|\widehat{\tau} - \tau_0|, \tag{E.39}$$

w.p.a.1. Combining (E.37)-(E.39) with a sufficiently small $\widetilde{c}$ yields

$$R(\widehat{\alpha}, \widehat{\tau}) + \omega_n \left| \widehat{D}(\widehat{\alpha} - \alpha_0) \right|_1 \leq C \left(\omega_n^2 s + |\widehat{\tau} - \tau_0| \right) \tag{E.40}$$

for some finite constant $C > 0$. Since $|\widehat{\tau} - \tau_0| = O_P(n^{-1})$ by Theorem 4.3, the desired results follow (E.40) immediately.

Now we consider case (i). In this case,

$$R(\widehat{\alpha}, \widehat{\tau}) + \frac{1}{2}\omega_n \left| \widehat{D}(\widehat{\alpha} - \alpha_0) \right|_1 \leq 3 \left| R(\alpha_0, \widehat{\tau}) \right|. \tag{E.41}$$

As shown in the proof of Theorem C.2, we have that

$$\left| R(\alpha_0, \widehat{\tau}) \right| = O_P\left(|\delta_0|_1 \, n^{-1} \log n \right) = O_P\left(\omega_n^2 s \right). \tag{E.42}$$

Therefore, we obtain the desired results in case (i) as well by combining (E.42) with (E.41).

E.8   Proof of Theorems 4.5

We write α_J be a subvector of α whose components' indices are in $J(\alpha_0)$. Define $\bar{Q}_n(\alpha_J) \equiv \widetilde{S}_n((\alpha_J, 0))$, so that

$$\bar{Q}_n(\alpha_J) = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_{iJ}(\widehat{\tau})^T \alpha_J) + \mu_n \sum_{j \in J(\alpha_0)} w_j \widehat{D}_j |\alpha_j|.$$

For notational simplicity, here we write $\widehat{D}_j \equiv D_j(\widehat{\tau})$. When τ_0 is identifiable, our argument is conditional on

$$\widehat{\tau} \in \mathcal{T}_n = \left\{ |\tau - \tau_0| \le n^{-1} \log n \right\}, \tag{E.43}$$

whose probability goes to 1 due to Theorem 4.3.

We first prove the following two lemmas. Define

$$\bar{\alpha}_J \equiv \underset{\alpha_J}{\operatorname{argmin}} \, \bar{Q}_n(\alpha_J). \tag{E.44}$$

Lemma E.4. *Suppose that $M_n^2 (\log n)^2 / (s \log s) = o(n)$, $s^4 \log s = o(n)$, $s^2 \log n / \log s = o(n)$ and $\widehat{\tau} \in \mathcal{T}_n$ if $\delta_0 \ne 0$; suppose that $s^4 \log s = o(n)$ and $\widehat{\tau}$ is any value in $\mathcal{T}$ if $\delta_0 = 0$. Then*

$$|\bar{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}} \right).$$

Proof of Lemma E.4. Let $k_n = \sqrt{\frac{s \log s}{n}}$. We first prove that for any $\epsilon > 0$, there is $C_\epsilon > 0$, with probability at least $1 - \epsilon$,

$$\inf_{|\alpha_J - \alpha_{0J}|_2 = C_\epsilon k_n} \bar{Q}_n(\alpha_J) > \bar{Q}_n(\alpha_{0J}) \tag{E.45}$$

Once this is proved, then by the continuity of $\bar{Q}_n$, there is a local minimizer of $\bar{Q}_n(\alpha_J)$ inside $B(\alpha_{0J}, C_\epsilon k_n) \equiv \{\alpha_J \in \mathbb{R}^s : |\alpha_{0J} - \alpha_J|_2 \le C_\epsilon k_n\}$. Due to the convexity of $\bar{Q}_n$, such a local minimizer is also global. We now prove (E.45).

Write

$$l_J(\alpha_J) = \frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_{iJ}(\widehat{\tau})^T \alpha_J), \quad L_J(\alpha_J, \tau) = \mathbb{E}[\rho(Y, X_J(\tau)^T \alpha_J)].$$

Then for all $|\alpha_J - \alpha_{0J}|_2 = C_\epsilon k_n$,

$$
\begin{aligned}
& \bar{Q}_n(\alpha_J) - \bar{Q}_n(\alpha_{0J}) \\
=\ & l_J(\alpha_J) - l_J(\alpha_{0J}) + \sum_{j \in J(\alpha_0)} w_j \mu_n \widehat{D}_j (|\alpha_j| - |\alpha_{0j}|) \\
\geq\ & \underbrace{L_J(\alpha_J, \widehat{\tau}) - L_J(\alpha_{0J}, \widehat{\tau})}_{(1)} - \underbrace{\sup_{|\alpha_J - \alpha_{0J}|_2 \leq C_\delta k_n} |\nu_n(\alpha_J, \widehat{\tau}) - \nu_n(\alpha_{0J}, \widehat{\tau})|}_{(2)} + \underbrace{\sum_{j \in J(\alpha_0)} \mu_n \widehat{D}_j w_j (|\alpha_j| - |\alpha_{0j}|)}_{(3)}.
\end{aligned}
$$

To analyze (1), note that $|\alpha_J - \alpha_{0J}|_2 = C_\epsilon k_n$ and $m_J(\tau_0, \alpha_0) = 0$ and when $\delta_0 = 0$, $m_J(\tau, \alpha_{0J})$ is free of τ. Then there is $c_3 > 0$,

$$
\begin{aligned}
& L_J(\alpha_J, \widehat{\tau}) - L_J(\alpha_{0J}, \widehat{\tau}) \\
\geq\ & m_J(\tau_0, \alpha_{0J})^T (\alpha_J - \alpha_{0J}) + (\alpha_J - \alpha_{0J})^T \frac{\partial^2 \mathbb{E}[\rho(Y, X_J(\widehat{\tau})^T \alpha_{0J})]}{\partial \alpha_J \partial \alpha_J^T} (\alpha_J - \alpha_{0J}) \\
& - |m_J(\tau_0, \alpha_{0J}) - m_J(\widehat{\tau}, \alpha_{0J})|_2 |\alpha_J - \alpha_{0J}|_2 - c_3 |\alpha_{0J} - \alpha_J|_1^3 \\
\geq\ & \lambda_{\min}\left(\frac{\partial^2 \mathbb{E}[\rho(Y, X_J(\widehat{\tau})^T \alpha_{0J})]}{\partial \alpha_J \partial \alpha_J^T} \right) |\alpha_J - \alpha_{0J}|_2^2 \\
& - (|m_J(\tau_0, \alpha_{0J}) - m_J(\widehat{\tau}, \alpha_{0J})|_2)|\alpha_J - \alpha_{0J}|_2 - c_3 s^{3/2} |\alpha_{0J} - \alpha_J|_2^3 \\
\geq\ & c_1 C_\epsilon^2 k_n^2 - (|m_J(\tau_0, \alpha_{0J}) - m_J(\widehat{\tau}, \alpha_{0J})|_2) C_\epsilon k_n - c_3 s^{3/2} C_\delta^3 k_n^3 \\
\geq\ & C_\epsilon k_n (c_1 C_\epsilon k_n - M_n n^{-1} \log n - c_3 s^{3/2} C_\epsilon^2 k_n^2) \geq c_1 C_\delta^2 k_n^2/3,
\end{aligned}
$$

where the last inequality follows from $M_n n^{-1} \log n < 1/3 c_1 C_\epsilon k_n$ and $c_3 s^{3/2} C_\epsilon^2 k_n^2 < 1/3 c_1 C_\epsilon k_n$. These follow from the conditions $M_n^2 (\log n)^2/(s \log s) = o(n)$ and $s^4 \log s = o(n)$.

To analyze (2), by the symmetrization theorem and the contraction theorem (see, for example, Theorems 14.3 and 14.4 of Bühlmann and van de Geer (2011)), there is a Rademacher sequence $\epsilon_1, ..., \epsilon_n$ independent of $\{Y_i, X_i, Q_i\}_{i \leq n}$ such that (note that when $\delta_0 = 0$, $\alpha_J = \beta_J$,

$$
\nu_n(\alpha_J, \tau) \equiv \frac{1}{n} \sum_{i=1}^n \left[\rho\left(Y_i, X_{J(\beta_0)i}^T \beta_J\right) - \mathbb{E}\rho\left(Y, X_{J(\beta_0)}^T \beta_J\right) \right],
$$

which is free of τ)

$$
\begin{aligned}
V_n &= \mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_2\leq C_\epsilon k_n}|\nu_n(\alpha_J,\tau)-\nu_n(\alpha_{0J},\tau)|\right) \\
&\leq 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_2\leq C_\epsilon k_n}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i[\rho(Y_i,X_{iJ}(\tau)^T\alpha_J)-\rho(Y_i,X_{iJ}(\tau)^T\alpha_{0J})]\right|\right) \\
&\leq 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_2\leq C_\epsilon k_n}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(X_{iJ}(\tau)^T(\alpha_J-\alpha_{0J}))\right|\right),
\end{aligned}
$$

which is bounded by the sum of the following two terms, $V_{1n}+V_{2n}$, due to the triangle inequality and the fact that $|\alpha_J-\alpha_{0J}|_1\leq|\alpha_J-\alpha_{0J}|_2\sqrt{s}$: first, when $\delta_0=0$, $V_{1n}\equiv0$; second, when $\delta_0\neq0$ and τ_0 is identifiable, we have that

$$
\begin{aligned}
V_{1n} &= 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\alpha_J-\alpha_{0J}|_1\leq C_\epsilon k_n\sqrt{s}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(X_{iJ}(\tau)-X_{iJ}(\tau_0))^T(\alpha_J-\alpha_{0J})\right|\right) \\
&\leq 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\sup_{|\delta_J-\delta_{0J}|_1\leq C_\epsilon k_n\sqrt{s}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{iJ(\delta_0)}^T(1\{Q_i>\tau\}-1\{Q_i>\tau_0\})(\delta_J-\delta_{0J})\right|\right) \\
&\leq 4LC_\epsilon k_n\sqrt{s}\,\mathbb{E}\left(\sup_{\tau\in\mathcal{T}_n}\max_{j\in J(\delta_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{ij}(1\{Q_i>\tau\}-1\{Q_i>\tau_0\})\right|\right) \\
&\leq 4LC_\epsilon k_n\sqrt{s}\,C_1\,|J(\delta_0)|_0\sqrt{\frac{\log n}{n^2}}
\end{aligned}
$$

by bounding the maximum over j with summation and using the maximal inequality in Theorem 2.14.1 in van der Vaart and Wellner (1996) since the class of transformations $\epsilon_i X_{ij}(1\{Q_i>\tau\}-1\{Q_i>\tau_0\})$ constitutes a VC class of functions. Here the bound is uniform and determined by the L_2-norm of the envelope, which is proportional to

$$
\sqrt{\mathbb{E}\left(1\{|Q_i-\tau_0|\leq n^{-1}\log n\}\right)}.
$$

Note that

$$
\begin{aligned}
V_{2n} &= 4L\mathbb{E}\left(\sup_{|\alpha_J-\alpha_{0J}|_1\le C_\epsilon k_n\sqrt{s}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{iJ}(\tau_0)^T(\alpha_J-\alpha_{0J})\right|\right)\\
&\le 4LC_\epsilon k_n\sqrt{s}\,\mathbb{E}\left(\max_{j\in J(\alpha_0)}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}(\tau_0)\right|\right)\le 4LC_\epsilon C_2 k_n^2,
\end{aligned}
$$

due to the Bernstein's moment inequality (Lemma 14.12 of Bühlmann and van de Geer (2011) for some $C_2>0$. Therefore,

$$
V_n\le 4LC_\epsilon k_n\sqrt{s}\,C_1\,|J(\delta_0)|_0\,\sqrt{\frac{\log n}{n^2}}+4LC_\epsilon C_2 k_n^2<5LC_\epsilon C_2 k_n^2,
$$

where the last inequality is due to $s^2\log n/\log s=o(n)$. Therefore, conditioning on the event $\widehat{\tau}\in\mathcal{T}_n$ when $\delta_0\ne 0$, or for $\widehat{\tau}\in\mathcal{T}$ when $\delta_0=0$, with probability at least $1-\epsilon$, $(2)\le\frac{1}{\epsilon}5LC_2C_\epsilon k_n^2$.

In addition, note that $P(\max_{j\in J(\alpha_0)}|w_j|=0)=1$, so $(3)=0$ with probability approaching one. Hence

$$
\inf_{|\alpha_J-\alpha_{0J}|_2=C_\epsilon k_n}\bar{Q}_n(\alpha_J)-\bar{Q}_n(\alpha_{0J})\ge\frac{c_1 C_\epsilon^2 k_n^2}{3}-\frac{1}{\epsilon}5LC_2C_\epsilon k_n^2>0.
$$

The last inequality holds for $C_\epsilon>\frac{15LC_2}{c_1\epsilon}$. By the continuity of $\bar{Q}_n$, there is a local minimizer of $\bar{Q}_n(\alpha_J)$ inside $\{\alpha_J\in\mathbb{R}^s:|\alpha_{0J}-\alpha_J|_2\le C_\epsilon k_n\}$, which is also a global minimizer due to the convexity. ∎

On $\mathbb{R}^{2p}$, recall that

$$
R_n(\tau,\alpha)=\frac{1}{n}\sum_{i=1}^n\rho(Y_i,X_i(\tau)^T\alpha).
$$

For $\bar{\alpha}_J=(\bar{\beta}_{J(\beta_0)},\bar{\delta}_{J(\delta_0)})\equiv(\bar{\beta}_J,\bar{\delta}_J)$ in the previous lemma, define

$$
\bar{\alpha}=(\bar{\beta}_J^T,0^T,\bar{\delta}_J^T,0^T)^T.
$$

Without introducing confusions, we also write $\bar{\alpha} = (\bar{\alpha}_J, 0)$ for notational simplicity. This notation indicates that $\bar{\alpha}$ has zero entries on the indices outside the oracle index set $J(\alpha_0)$. We prove the following lemma.

Lemma E.5. *With probability approaching one, there is a random neighborhood of $\bar{\alpha}$ in $\mathbb{R}^{2p}$, denoted by $\mathcal{H}$, so that $\forall \alpha = (\alpha_J, \alpha_{J^c}) \in \mathcal{H}$, if $\alpha_{J^c} \neq 0$, we have $\widetilde{S}_n(\alpha_J, 0) < \widetilde{Q}_n(\alpha)$.*

Proof of Lemma E.5. Define an l_2-ball, for $r_n \equiv \mu_n / \log n$,

$$\mathcal{H} = \{\alpha \in \mathbb{R}^{2p} : |\alpha - \bar{\alpha}|_2 < r_n/(2p)\}.$$

Then $\sup_{\alpha \in \mathcal{H}} |\alpha - \bar{\alpha}|_1 = \sup_{\alpha \in \mathcal{H}} \sum_{l \leq 2p} |\alpha_l - \bar{\alpha}_l| < r_n$. Consider any $\tau \in \mathcal{T}_n$. For any $\alpha = (\alpha_J, \alpha_{J^c}) \in \mathcal{H}$, write

$$R_n(\tau, \alpha_J, 0) - R_n(\tau, \alpha)$$
$$= R_n(\tau, \alpha_J, 0) - \mathbb{E}R_n(\tau, \alpha_J, 0) + \mathbb{E}R_n(\tau, \alpha_J, 0) - R_n(\tau, \alpha) + \mathbb{E}R_n(\tau, \alpha) - \mathbb{E}R_n(\tau, \alpha)$$
$$\leq \mathbb{E}R_n(\tau, \alpha_J, 0) - \mathbb{E}R_n(\tau, \alpha) + |R_n(\tau, \alpha_J, 0) - \mathbb{E}R_n(\tau, \alpha_J, 0) + \mathbb{E}R_n(\tau, \alpha) - R_n(\tau, \alpha)|$$
$$\leq \mathbb{E}R_n(\tau, \alpha_J, 0) - \mathbb{E}R_n(\tau, \alpha) + |\nu_n(\alpha_J, 0, \tau) - \nu_n(\alpha, \tau)|.$$

Note that $|(\alpha_J, 0) - \bar{\alpha}|_2^2 = |\alpha_J - \bar{\alpha}_J|_2^2 \leq |\alpha_J - \bar{\alpha}_J|_2^2 + |\alpha_{J^c} - 0|_2^2 = |\alpha - \bar{\alpha}|_2^2$. Hence $\alpha \in \mathcal{H}$ implies $(\alpha_J, 0) \in \mathcal{H}$. In addition, by definition of $\bar{\alpha} = (\bar{\alpha}_J, 0)$ and $|\bar{\alpha}_J - \alpha_{0J}|_2 = O_P(\sqrt{\frac{s \log s}{n}})$ (Lemma E.4), we have $|\bar{\alpha} - \alpha_0|_1 = O_P(s\sqrt{\frac{\log s}{n}})$, which also implies

$$\sup_{\alpha \in \mathcal{H}} |\alpha - \alpha_0|_1 = O_P\left(s\sqrt{\frac{\log s}{n}}\right) + r_n,$$

where the randomness in $\sup_{\alpha \in \mathcal{H}} |\alpha - \alpha_0|_1$ comes from that of $\mathcal{H}$.

By the mean value theorem, there is h in the segment between α and $(\alpha_J, 0)$,

$$\mathbb{E}R_n(\tau, \alpha_J, 0) - \mathbb{E}R_n(\tau, \alpha) = \mathbb{E}\rho(Y, X_J(\tau)^T \alpha_J) - \mathbb{E}\rho(Y, X_J(\tau)^T \alpha_J + X_{J^c}(\tau)^T \alpha_{J^c})$$

$$= -\sum_{j \notin J(\alpha_0)} \frac{\partial \mathbb{E}\rho(Y, X(\tau)^T h)}{\partial \alpha_j} \alpha_j \equiv \sum_{j \notin J(\alpha_0)} m_j(\tau, h) \alpha_j$$

where $m_j(\tau, h) = -\frac{\partial \mathbb{E}\rho(Y, X(\tau)^T h)}{\partial \alpha_j}$. Hence, $\mathbb{E}R_n(\tau, \alpha_J, 0) - \mathbb{E}R_n(\tau, \alpha) \leq \sum_{j \notin J} |m_j(\tau, h)| |\alpha_j|$.

Because h is on the segment between α and $(\alpha_J, 0)$, so $h \in \mathcal{H}$. So for all $j \notin J(\alpha_0)$,

$$|m_j(\tau, h)| \leq \sup_{\alpha \in \mathcal{H}} |m_j(\tau, \alpha)| \leq \sup_{\alpha \in \mathcal{H}} |m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| + |m_j(\tau, \alpha_0) - m_j(\tau_0, \alpha_0)|.$$

We now argue that we can apply Assumption 8 (ii). Let

$$c_n \equiv s\sqrt{(\log s)/n} + r_n.$$

For any $\epsilon > 0$, there is $C_\epsilon > 0$, with probability at last $1 - \epsilon$, $\sup_{\alpha \in \mathcal{H}} |\alpha - \alpha_0|_1 \leq C_\epsilon c_n$. $\forall \alpha \in \mathcal{H}$, write $\alpha = (\beta, \delta)$ and $\theta = \beta + \delta$. On the event $|\alpha - \alpha_0|_1 \leq C_\epsilon c_n$, we have $|\beta - \beta_0|_1 \leq C_\epsilon c_n$ and $|\theta - \theta_0|_1 \leq C_\epsilon c_n$. Hence $\mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau_0\}] \leq |\beta - \beta_0|_1^2 \max_{i,j \leq p} E|X_i X_j| < r^2$, yielding $\beta \in \mathcal{B}(\beta_0, r)$. Similarly, $\theta \in \mathcal{G}(\theta_0, r)$. Therefore, by Assumption 8 (ii), with probability at least $1 - \epsilon$, (note that neither C_ϵ, L nor c_n depend on α)

$$\max_{j \notin J(\alpha_0)} \sup_{\tau \in \mathcal{T}_n} \sup_{\alpha \in \mathcal{H}} |m_j(\tau, \alpha) - m_j(\tau, \alpha_0)| \leq L \sup_{\alpha \in \mathcal{H}} |\alpha - \alpha_0|_1 \leq L(C_\epsilon c_n),$$

$$\max_{j \leq 2p} \sup_{\tau \in \mathcal{T}_n} |m_j(\tau, \alpha_0) - m_j(\tau_0, \alpha_0)| \leq M_n n^{-1} \log n.$$

In particular, when $\delta_0 = 0$, $m_j(\tau, \alpha_0) = 0$ for all τ. Therefore, when $\delta_0 \neq 0$,

$$\sup_{j \notin J(\alpha_0)} \sup_{\tau \in \mathcal{T}_n} |m_j(\tau, h)| = O_P(c_n + M_n n^{-1} \log n) = o_P(\mu_n);$$

when $\delta_0 = 0$, $\sup_{j \notin J(\alpha_0)} \sup_{\tau \in \mathcal{T}} |m_j(\tau, h)| = O_P(c_n) = o_P(\mu_n)$.

Let $\epsilon_1, ..., \epsilon_n$ be a Rademacher sequence independent of $\{Y_i, X_i, Q_i\}_{i \leq n}$. Then by the

symmetrization and contraction theorems,

$$
\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}|\nu_n(\alpha_J,0,\tau)-\nu_n(\alpha,\tau)|\right)
$$

$$
\leq 2\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i[\rho(Y_i,X_{iJ}(\tau)^T\alpha_J)-\rho(Y_i,X_i(\tau)^T\alpha)]\right|\right)
$$

$$
\leq 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i[X_{iJ}(\tau)^T\alpha_J-X_i(\tau)^T\alpha]\right|\right)
$$

$$
\leq 4L\mathbb{E}\left(\sup_{\tau\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_iX_i(\tau)\right\|_{\max}\right)\sum_{j\notin J(\alpha_0)}|\alpha_j|\leq 2\omega_n\sum_{j\notin J(\alpha_0)}|\alpha_j|,
$$

where the last equality follows from (E.5).

Thus uniformly over $\alpha\in\mathcal{H}$, $R_n(\tau,\alpha_J,0)-R_n(\tau,\alpha)=o_P(\mu_n)\sum_{j\notin J(\alpha_0)}|\alpha_j|$. On the other hand,

$$
\sum_{j\in J(\alpha_0)}w_j\mu_n\widehat{D}_j|\alpha_j|-\sum_{j}w_j\mu_n\widehat{D}_j|\alpha_j|=\sum_{j\notin J(\alpha_0)}\mu_nw_j\widehat{D}_j|\alpha_j|.
$$

Also, w.p.a.1, $w_j=1$ and $\widehat{D}_j\geq\overline{D}$ for all $j\notin J(\alpha_0)$. Hence with probability approaching one, $\widetilde{Q}_n(\alpha_J,0)-\widetilde{Q}_n(\alpha)$ equals

$$
R_n(\widehat{\tau},\alpha_J,0)+\sum_{j\in J(\alpha_0)}\widehat{D}_jw_j\lambda_n|\alpha_j|-R_n(\widehat{\tau},\alpha)-\sum_{j\leq 2p}\widehat{D}_jw_j\omega_n|\alpha_j|\leq-\underline{D}\frac{\mu_n}{2}\sum_{j\notin J(\alpha_0)}|\alpha_j|<0. \ \blacksquare
$$

Proof of Theorem 4.5. Conditions in Lemmas E.4 and E.5 are expressed in terms of M_n. By Lemma D.1, we verify that in quantile regression models, $M_n=Cs^{1/2}$ for some $C>0$. Then all the required conditions in Lemmas E.4 and E.5 are satisfied by the conditions imposed in Theorem 4.5.

By Lemmas E.4 and E.5, w.p.a.1, for any $\alpha=(\alpha_J,\alpha_{J^c})\in\mathcal{H}$,

$$
\widetilde{S}_n(\bar{\alpha}_J,0)=\bar{Q}_n(\bar{\alpha}_J)\leq\bar{Q}_n(\alpha_J)=\widetilde{S}_n(\alpha_J,0)\leq\widetilde{S}_n(\alpha).
$$

Hence $(\bar{\alpha}_J,0)$ is a local minimizer of $\widetilde{S}_n$, which is also a global minimizer due to the convexity.

This implies that w.p.a.1, $\widetilde{\alpha} = (\widetilde{\alpha}_J, \widetilde{\alpha}_{J^c})$ satisfies: $\widetilde{\alpha}_{J^c} = 0$, and $\widetilde{\alpha}_J = \bar{\alpha}_J$, so

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right), \quad |\widetilde{\alpha}_J - \alpha_{0J}|_1 = O_P\left(s\sqrt{\frac{\log s}{n}}\right).$$

Finally, by (E.48), and that $R(\alpha_0, \widehat{\tau}) \le Cs|\widehat{\tau} - \tau_0| = O_P(sn^{-1})$,

$$R(\widetilde{\alpha}, \widehat{\tau}) \le 2R(\alpha_0, \widehat{\tau}) + 3\mu_n \bar{D}|\widetilde{\alpha} - \alpha_0|_1 = O_P(sn^{-1} + \mu_n s\sqrt{\frac{\log s}{n}}) = O_P(\mu_n s\sqrt{\frac{\log s}{n}}).$$

∎

E.9   Proof of Theorem 4.6

Recall that by Theorems 4.3 and 4.5, we have

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s \log s}{n}}\right) \quad \text{and} \quad |\widehat{\tau} - \tau_0| = O_P(n^{-1}), \tag{E.46}$$

and the set of regressors with nonzero coefficients is recovered w.p.a.1. Hence we can restrict ourselves on the oracle space $J(\alpha_0)$. In view of (E.46), define $r_n \equiv \sqrt{n^{-1}s \log s}$ and s_n. Let

$$R_n^*(\alpha_J, \tau) \equiv \frac{1}{n}\sum_{i=1}^n \rho\left(Y_i, X_{iJ}(\tau)^T \alpha_J\right),$$

where $\alpha_J \in \mathcal{A}_n \equiv \{\alpha_J : |\alpha_J - \alpha_{0J}|_2 \le Kr_n\} \subset \mathbb{R}^s$ and $\tau \in \mathcal{T}_n \equiv \{\tau : |\tau - \tau_0| \le Ks_n\}$ for some $K < \infty$, where K is a generic finite constant.

The following lemma is useful to establish that α_0 can be estimated as if τ_0 were known.

Lemma E.6 (Asymptotic Equivalence). *Assume that $\frac{\partial}{\partial \alpha}E\left[\rho\left(Y, X^T\alpha\right)|Q = t\right]$ exists for all t in a neighborhood of τ_0 and all its elements are continuous and bounded. Suppose that $s^3(\log s)(\log n) = o(n)$. Then*

$$\sup_{\alpha_J \in \mathcal{A}_n, \tau \in \mathcal{T}_n} |\{R_n^*(\alpha_J, \tau) - R_n^*(\alpha_J, \tau_0)\} - \{R_n^*(\alpha_{0J}, \tau) - R_n^*(\alpha_{0J}, \tau_0)\}| = o_P\left(n^{-1}\right).$$

This lemma implies that the asymptotic distribution of $\text{argmin}_{\alpha_J} R_n^* (\alpha_J, \widehat{\tau})$ can be characterized by $\widehat{\alpha}_J^* \equiv \text{argmin}_{\alpha_J} R_n^* (\alpha_J, \tau_0)$. It then follows immediately from the variable selection consistency that the asymptotic distribution of $\widetilde{\alpha}_J$ is equivalent to that of $\widehat{\alpha}_J^*$. Therefore, we have proved the theorem.

Proof of Lemma E.6. Noting that

$$\rho \left(Y_i, X_i^T \beta + X_i^T \delta 1 \{Q_i > \tau\} \right) = \rho \left(Y_i, X_i^T \beta \right) 1 \{Q_i \le \tau\} + \rho \left(Y_i, X_i^T \beta + X_i^T \delta \right) 1 \{Q_i > \tau\},$$

we have, for $\tau > \tau_0$,

$$
\begin{aligned}
D_n &(\alpha, \tau) \\
&\equiv \{R_n (\alpha, \tau) - R_n (\alpha, \tau_0)\} - \{R_n (\alpha_0, \tau) - R_n (\alpha_0, \tau_0)\} \\
&= \frac{1}{n} \sum_{i=1}^n \left[\rho \left(Y_i, X_i^T \beta \right) - \rho \left(Y_i, X_i^T \beta_0 \right) \right] 1 \{\tau_0 < Q_i \le \tau\} \\
&\quad - \frac{1}{n} \sum_{i=1}^n \left[\rho \left(Y_i, X_i^T \beta + X_i^T \delta \right) - \rho \left(Y_i, X_i^T \beta_0 + X_i^T \delta_0 \right) \right] 1 \{\tau_0 < Q_i \le \tau\} \\
&=: D_{n1} (\alpha, \tau) - D_{n2} (\alpha, \tau).
\end{aligned}
$$

To prove this lemma, we consider empirical processes

$$\mathbb{G}_{nj} (\alpha_J, \tau) \equiv \sqrt{n} \left(D_{nj} (\alpha_J, \tau) - \mathbb{E} D_{nj} (\alpha_J, \tau) \right), \quad (j = 1, 2),$$

and apply the maximal inequality in Theorem 2.14.2 of van der Vaart and Wellner (1996).

First, for $\mathbb{G}_{n1} (\alpha_J, \tau)$, we consider the following class of functions indexed by (β_J, τ):

$$\mathcal{F}_n \equiv \{\left(\rho \left(Y_i, X_{iJ}^T \beta_J \right) - \rho \left(Y_i, X_{iJ}^T \beta_{0J} \right) \right) 1 \left(\tau_0 < Q_i \le \tau \right) : |\beta_J - \beta_{0J}|_2 \le K r_n \text{ and } |\tau - \tau_0| \le K s_n\}.$$

Note that the Lipschitz property of ρ yields that

$$\left|\rho\left(Y_i, X_{iJ}^T\beta_J\right) - \rho\left(Y_i, X_{iJ}^T\beta_{0J}\right)\right| 1\left\{\tau_0 < Q_i \leq \tau\right\} \leq \left|X_{iJ}^T\right|_2 |\beta_J - \beta_{0J}|_2 1\left\{|Q_i - \tau_0| \leq Ks_n\right\}.$$

Thus, we let the envelope function be $F_n(X_{iJ}, Q_i) \equiv |X_{iJ}|_2 Kr_n 1\left\{|Q_i - \tau_0| \leq Ks_n\right\}$ and note that its L_2 norm is $O\left(\sqrt{s}r_n\sqrt{s_n}\right)$.

To compute the bracketing integral

$$J_{[]}\left(1, \mathcal{F}_n, L_2\right) \equiv \int_0^1 \sqrt{1 + \log N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_n, L_2\right)}d\varepsilon,$$

note that its 2ε bracketing number is bounded by the product of the ε bracketing numbers of two classes $\mathcal{F}_{n1} \equiv \left\{\rho\left(Y_i, X_{iJ}^T\beta_J\right) - \rho\left(Y_i, X_{iJ}^T\beta_0\right) : |\beta_J - \beta_{0J}|_2 \leq Kr_n\right\}$ and $\mathcal{F}_{n2} \equiv \left\{1\left(\tau_0 < Q_i \leq \tau\right) : |\tau - \tau_0| \leq Ks_n\right\}$ by Lemma 9.25 of Kosorok (2008) since both classes are bounded w.p.a.1 (note that w.p.a.1, $|X_{iJ}|_2 Kr_n < C < \infty$ for some constant C). That is,

$$N_{[]}\left(2\varepsilon\|F_n\|_{L_2}, \mathcal{F}_n, L_2\right) \leq N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n1}, L_2\right) N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n2}, L_2\right).$$

Let $F_{n1}(X_{iJ}) \equiv |X_{iJ}|_2 Kr_n$ and $l_n(X_{iJ}) \equiv |X_{iJ}|_2$. Note that by Theorem 2.7.11 of van der Vaart and Wellner (1996), the Lipschitz property of ρ implies that

$$N_{[]}\left(2\varepsilon\|l_n\|_{L_2}, \mathcal{F}_{n1}, L_2\right) \leq N(\varepsilon, \{\beta_J : |\beta_J - \beta_{0J}|_2 \leq Kr_n\}, |\cdot|_2),$$

which in turn implies that, for some constant C,

$$\begin{aligned}
N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n1}, L_2\right) &\leq N\left(\frac{\varepsilon\|F_n\|_{L_2}}{2\|l_n\|_{L_2}}, \{\beta_J : |\beta_J - \beta_{0J}|_2 \leq Kr_n\}, |\cdot|_2\right)\\
&\leq C\left(\frac{\sqrt{s}}{\varepsilon\sqrt{s_n}}\right)^s = C\left(\frac{\sqrt{ns}}{\varepsilon}\right)^s,
\end{aligned}$$

where the last inequality holds since a ε-ball contains a hypercube with side length $\varepsilon/\sqrt{s}$ in

the s-dimensional Euclidean space. On the other hand, for the second class of functions $\mathcal{F}_{n2}$ with the envelope function $F_{n2}(Q_i) \equiv 1\{|Q_i - \tau_0| \le Ks_n\}$, we have that

$$N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_{n2}, L_2\right) \le C\frac{\sqrt{s_n}}{\varepsilon\|F_n\|_{L_2}} = \frac{C}{\varepsilon\sqrt{s}r_n} = \frac{C\sqrt{n}}{\varepsilon s\sqrt{\log s}},$$

for some constant C. Combining these results together yields that

$$N_{[]}\left(\varepsilon\|F_n\|_{L_2}, \mathcal{F}_n, L_2\right) \le \frac{C^2\sqrt{n}}{\varepsilon s\sqrt{\log s}}\left(\frac{\sqrt{ns}}{\varepsilon}\right)^s \le C^2\varepsilon^{-s-1}n^{(s+1)/2}$$

for all sufficiently large n. Then we have that

$$J_{[]}\left(1, \mathcal{F}_n, L_2\right) \le C^2(\sqrt{s\log n} + \sqrt{s})$$

for all sufficiently large n. Thus, by the maximal inequality in Theorem 2.14.2 of van der Vaart and Wellner (1996),

$$
\begin{aligned}
n^{-1/2} \mathbb{E} \sup_{\mathcal{A}_n \times \mathcal{T}_n} |\mathbb{G}_{n1}\left(\alpha_J, \tau\right)| &\le O\left[n^{-1/2}\sqrt{s}r_n\sqrt{s_n}(\sqrt{s\log n} + \sqrt{s})\right] \\
&= O\left[\frac{s}{n^{3/2}}\sqrt{\log s}(\sqrt{s\log n} + \sqrt{s})\right] \\
&= o\left(n^{-1}\right),
\end{aligned}
$$

where the last equality follows from the restriction that $s^3(\log s)(\log n) = o(n)$. Identical arguments also apply to $\mathbb{G}_{n2}\left(\alpha_J, \tau\right)$.

Turning to $\mathbb{E}D_n\left(\alpha, \tau\right)$, note that by the condition that $\frac{\partial}{\partial \alpha}E\left[\rho\left(Y, X^T\alpha\right)|Q = t\right]$ exists for all t in a neighborhood of τ_0 and all its elements are continuous and bounded, we have

that for some mean value $\tilde{\beta}_J$ between β_J and β_{0J},

$$
\begin{aligned}
\left| \mathbb{E} \left(\rho \left(Y, X_J^T \beta_J \right) - \rho \left(Y, X_J^T \beta_{0J} \right) \right) 1 \left\{ \tau_0 < Q \le \tau \right\} \right| \\
= \left| \mathbb{E} \left[\frac{\partial}{\partial \beta} \mathbb{E} \left[\rho \left(Y, X^T \tilde{\beta}_J \right) | Q \right] 1 \left\{ \tau_0 < Q \le \tau \right\} \right] (\beta - \beta_0) \right| \\
= O \left(s r_n s_n \right) \\
= O \left[\frac{s^{3/2}}{n^{3/2}} \sqrt{\log s} \right] \\
= o \left(n^{-1} \right),
\end{aligned}
$$

where the last equality follows from the restriction that $s^3 (\log s) = o(n)$. Since the same holds for the other term in $\mathbb{E} D_n$, $\sup |\mathbb{E} D_n (\alpha, \tau)| = o(n^{-1})$ as desired. ∎

E.10 Proof of Theorem 4.7

By definition,

$$
\frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_i(\widehat{\tau})^T \widetilde{\alpha}) + \mu_n |W \widehat{D} \widetilde{\alpha}|_1 \le \frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_i(\widehat{\tau})^T \alpha_0) + \mu_n |W \widehat{D} \alpha_0|_1.
$$

where $W = \text{diag}\{w_1, ..., w_{2p}\}$. From this, we obtain the following inequality

$$
R(\widetilde{\alpha}, \widehat{\tau}) + \mu_n |W \widehat{D} \widetilde{\alpha}|_1 \le |\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widetilde{\alpha}, \widehat{\tau})| + R(\alpha_0, \widehat{\tau}) + \mu_n |W \widehat{D} \alpha_0|_1.
$$

Now applying Lemma E.1 yields, when $\sqrt{\log(np)/n} = o(\mu_n)$ (which is true under the assumption that $\omega_n \ll \mu_n$), we have that w.p.a.1, $|\nu_n(\alpha_0, \widehat{\tau}) - \nu_n(\widetilde{\alpha}, \widehat{\tau})| \le \frac{1}{2} \mu_n |\widehat{D}(\alpha_0 - \widetilde{\alpha})|_1$. Hence on this event,

$$
R(\widetilde{\alpha}, \widehat{\tau}) + \mu_n |W \widehat{D} \widetilde{\alpha}|_1 \le \frac{1}{2} \mu_n |\widehat{D}(\alpha_0 - \widetilde{\alpha})|_1 + R(\alpha_0, \widehat{\tau}) + \mu_n |W \widehat{D} \alpha_0|_1.
$$

Note that $\max_j w_j \leq 1$, so for $\Delta := \widetilde{\alpha} - \alpha_0$,

$$R(\widetilde{\alpha}, \widehat{\tau}) + \mu_n |(W\widehat{D}\Delta)_{J^c}|_1 \leq \frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 + \frac{1}{2}\mu_n|\widehat{D}\Delta_{J^c}|_1 + R(\alpha_0, \widehat{\tau}).$$

By Theorem 4.2, $\max_{j \notin J} |\widehat{\alpha}_j| = O_P(\omega_n s)$. Hence for any $\epsilon > 0$, there is $C > 0$, $\max_{j \notin J} |\widehat{\alpha}_j| \leq C\omega_n s < \mu_n$ with probability at least $1 - \epsilon$. On the event $\max_{j \notin J} |\widehat{\alpha}_j| \leq C\omega_n s < \mu_n$, by definition, $w_j = 1 \; \forall j \notin J$. Hence on this event,

$$R(\widetilde{\alpha}, \widehat{\tau}) + \frac{1}{2}\mu_n|(\widehat{D}\Delta)_{J^c}|_1 \leq \frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 + R(\alpha_0, \widehat{\tau}). \qquad (E.47)$$

We now consider two cases: (i) $\frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 \leq R(\alpha_0, \widehat{\tau})$; (ii) $\frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 > R(\alpha_0, \widehat{\tau})$.

case 1: $\frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 \leq R(\alpha_0, \widehat{\tau})$

We have: for $C = 14\underline{D}^{-1}/3$, $\mu_n|\Delta|_1 \leq CR(\alpha_0, \widehat{\tau})$. If $\widehat{\tau} > \tau_0$, for $\tau = \widehat{\tau}$ in the inequalities below,

$$R(\alpha_0, \widehat{\tau}) = \mathbb{E}(\rho(Y, X^T\beta_0) - \rho(X^T\theta_0))1\{\tau_0 < Q < \tau\} \leq L\mathbb{E}|X^T\delta_0|1\{\tau_0 < Q < \tau\}$$
$$\leq L|\delta_0|_1 \max_{j \leq p} E|X_j|1\{\tau_0 < Q < \tau\} \leq L|\delta_0|_1 \max_{j \leq p} \sup_q E(|X_j||Q = q)P(\tau_0 < Q < \tau)$$
$$\leq Cs(\tau - \tau_0).$$

The case for $\tau \leq \tau_0$ follows from the same argument. Hence $\mu_n|\Delta|_1 \leq C|\widehat{\tau} - \tau_0|s$.

case 2: $\frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 > R(\alpha_0, \widehat{\tau})$

Then by the compatibility property,

$$R(\widetilde{\alpha}, \widehat{\tau}) + \frac{1}{2}\mu_n|(\widehat{D}\Delta)_{J^c}|_1 \leq 3\mu_n|\widehat{D}\Delta_J|_1 \leq 3\mu_n\bar{D}\sqrt{s}\|X(\tau_0)^T\Delta\|_2/\sqrt{\phi}.$$

The same argument as that of Step 5 in the proof of Theorem 4.2 yields

$$\|X(\tau_0)^T\Delta\|_2^2 \leq CR(\widetilde{\alpha}, \widehat{\tau}) + C|\widehat{\tau} - \tau_0|$$

for some generic constant $C > 0$. This implies, for some generic constant $C > 0$,

$$R(\widetilde{\alpha}, \widehat{\tau})^2 \leq \mu_n^2 sC(R(\widetilde{\alpha}, \widehat{\tau}) + |\widehat{\tau} - \tau_0|).$$

It follows that $R(\widetilde{\alpha}, \widehat{\tau}) \leq C(\mu_n^2 s + |\widehat{\tau} - \tau_0|)$, and $\|X(\tau_0)\Delta\|_2^2 \leq C(\mu_n^2 s + |\widehat{\tau} - \tau_0|)$. Hence

$$|\Delta|_1^2 \leq Cs\|X(\tau_0)\Delta\|_2^2 \leq C(\mu_n^2 s^2 + |\widehat{\tau} - \tau_0|s).$$

Combining both cases, we reach:

$$|\widetilde{\alpha} - \alpha_0|_1^2 \leq C(\mu_n^2 s^2 + |\widehat{\tau} - \tau_0|s + \frac{1}{\mu_n^2}|\widehat{\tau} - \tau_0|^2 s^2),$$

which gives the desired result since the first term $\mu_n^2 s^2$ dominates the other two terms.

Rate of convergence for $R(\tilde{\alpha}, \widehat{\tau})$

In the proofs above, we have in fact shown that

$$R(\widetilde{\alpha}, \widehat{\tau}) \leq 2R(\alpha_0, \widehat{\tau}) + 3\mu_n \bar{D}|\tilde{\alpha} - \alpha_0|_1, \tag{E.48}$$

and when $\delta_0 \neq 0$, $R(\alpha_0, \widehat{\tau}) \leq Cs|\widehat{\tau} - \tau_0|$. Note that $\widehat{\tau} - \tau_0 = O_P(n^{-1})$. Hence $R(\tilde{\alpha}, \widehat{\tau}) = O_P(sn^{-1} + \mu_n^2 s) = O_P(\mu_n^2 s)$.

E.11 Proof of Theorem 5.1

If $\delta_0 = 0$, τ_0 is non-identifiable. In this case, we decompose the excess risk in the following way:

$$
\begin{aligned}
R(\alpha, \tau) = {}& \mathbb{E}\left(\left[\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right]\mathbb{1}\left\{Q \leq \tau\right\}\right) \\
& + \mathbb{E}\left(\left[\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right]\mathbb{1}\left\{Q > \tau\right\}\right).
\end{aligned}
\tag{E.49}
$$

We split the proof into three steps.

Step 1: For any $r > 0$, we have that w.p.a.1, $\breve{\beta} \in \tilde{\mathcal{B}}(\beta_0, r, \breve{\tau})$ and $\breve{\theta} \in \tilde{\mathcal{G}}(\beta_0, r, \breve{\tau})$.

Proof of Step 1. As in the proof of Step 1 in the proof of Theorem 4.2, Assumption 9 (iii) implies that

$$\mathbb{E}\left[(X^T(\beta - \beta_0))^2 1\{Q \leq \tau\}\right] \leq \frac{R(\alpha, \tau)^2}{(\eta^* r^*)^2} \vee \frac{R(\alpha, \tau)}{\eta^*}.$$

For any $r > 0$, note that $R(\breve{\alpha}, \breve{\tau}) = o_P(1)$ implies that the event $R(\breve{\alpha}, \breve{\tau}) < r^2$ holds w.p.a.1. Therefore, we have shown that $\breve{\beta} \in \tilde{\mathcal{B}}(\beta_0, r, \breve{\tau})$. The other case can be proved similarly. ∎

Step 2 : Suppose that $\delta_0 = 0$. Then

$$R(\breve{\alpha}, \breve{\tau}) + \frac{1}{2}\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 \leq 2\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)_J\right|_1 \quad \text{w.p.a.1.} \tag{E.50}$$

Proof. The proof of this step is similar to that of Step 3 in the proof of Theorem 4.2. Since $(\breve{\alpha}, \breve{\tau})$ minimizes the ℓ_1-penalized objective function in (2.2), we have that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\breve{\tau})^T \breve{\alpha}) + \kappa_n |\breve{D}\breve{\alpha}|_1 \leq \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, X_i(\breve{\tau})^T \alpha_0) + \kappa_n |\breve{D}\alpha_0|_1. \tag{E.51}$$

When $\delta_0 = 0$, $\rho(Y, X(\breve{\tau})^T \alpha_0) = \rho(Y, X(\tau_0)^T \alpha_0)$. Using this fact and (E.51), we obtain the following inequality

$$R(\breve{\alpha}, \breve{\tau}) \leq [\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})] + \kappa_n |\breve{D}\alpha_0|_1 - \kappa_n |\breve{D}\breve{\alpha}|_1. \tag{E.52}$$

As in Step 3 in the proof of Theorem 4.2, we apply Lemma E.1 to $[\nu_n(\alpha_0, \breve{\tau}) - \nu_n(\breve{\alpha}, \breve{\tau})]$ with a_n and b_n replaced by $a_n/2$ and $b_n/2$. Then we can rewrite the basic inequality in (E.52) by

$$\kappa_n \left|\breve{D}\alpha_0\right|_1 \geq R(\breve{\alpha}, \breve{\tau}) + \kappa_n \left|\breve{D}\breve{\alpha}\right|_1 - \frac{1}{2}\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 \quad \text{w.p.a.1.}$$

Now adding $\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1$ on both sides of the inequality above and using the fact that

100

$|\alpha_{0j}|_1 - |\breve{\alpha}_j|_1 + |(\breve{\alpha}_j - \alpha_{0j})|_1 = 0$ for $j \notin J$, we have that w.p.a.1,

$$2\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)_J\right|_1 \geq R(\breve{\alpha}, \breve{\tau}) + \frac{1}{2}\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1.$$

Therefore, we have obtained the desired result. ∎

Step 3 : Suppose that $\delta_0 = 0$. Then

$$R(\breve{\alpha}, \breve{\tau}) = O_P(\kappa_n^2 s) \quad \text{and} \quad |\breve{\alpha} - \alpha_0| = O_P(\kappa_n s).$$

Proof. By Step 2,

$$4\left|\breve{D}(\breve{\alpha} - \alpha_0)_J\right|_1 \geq \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 = \left|\breve{D}(\breve{\alpha} - \alpha_0)_J\right|_1 + \left|\breve{D}(\breve{\alpha} - \alpha_0)_{J^c}\right|_1, \quad \text{(E.53)}$$

which enables us to apply the compatibility condition in Assumption 3.

Recall that $\|Z\|_2 = (EZ^2)^{1/2}$ for a random variable Z. Note that for $s = |J(\alpha_0)|_0$,

$$\begin{aligned}
&R(\breve{\alpha}, \breve{\tau}) + \frac{1}{2}\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)\right|_1 \\
&\leq_{(1)} 2\kappa_n \left|\breve{D}(\breve{\alpha} - \alpha_0)_J\right|_1 \\
&\leq_{(2)} 2\kappa_n \bar{D} \left\|X(\breve{\tau})^T(\breve{\alpha} - \alpha_0)\right\|_2 \sqrt{s}/\phi \\
&\leq_{(3)} \frac{4\kappa_n^2 \bar{D}^2 s}{2\tilde{c}\phi^2} + \frac{\tilde{c}}{2}\left\|X(\breve{\tau})^T(\breve{\alpha} - \alpha_0)\right\|_2^2,
\end{aligned} \quad \text{(E.54)}$$

where (1) is from the basic inequality (E.50) in Step 2, (2) is by the compatibility condition (Assumption 3), and (3) is from the inequality that $uv \leq v^2/(2\tilde{c}) + \tilde{c}u^2/2$ for any $\tilde{c} > 0$.

Note that

$$\left\|X(\tau)^T\alpha - X(\tau)^T\alpha_0\right\|_2^2$$

$$=_{(1)} \mathbb{E}\left[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}\right] + \mathbb{E}\left[(X^T(\beta - \beta_0))^2 1\{Q \le \tau\}\right]$$

$$\le_{(2)} (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\theta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q > \tau\}\right]$$

$$+ (\eta^*)^{-1}\mathbb{E}\left[\left(\rho\left(Y, X^T\beta\right) - \rho\left(Y, X^T\beta_0\right)\right) 1\{Q \le \tau\}\right]$$

$$\le_{(3)} (\eta^*)^{-1}R(\alpha, \tau), \tag{E.55}$$

where (1) is simply an identity, (2) from Assumption 9 (iii) , and (3) is due to (E.49). Hence, (E.54) with $\tilde{c} = \eta^*$ implies that

$$R\left(\breve{\alpha}, \breve{\tau}\right) + \kappa_n\left|\breve{D}\left(\breve{\alpha} - \alpha_0\right)\right|_1 \le \frac{4\kappa_n^2\bar{D}^2 s}{\eta^*\phi^2}. \tag{E.56}$$

Therefore, $R\left(\breve{\alpha}, \breve{\tau}\right) = O_P(\kappa_n^2 s)$. Also, $|\breve{\alpha} - \alpha_0| = O_P\left(\kappa_n s\right)$ since $D(\breve{\tau}) \ge \underline{D}$ w.p.a.1 by Assumption 1 (iv). $\blacksquare$

E.12   Proof of Theorem 5.2

We first prove part (i) when the minimum signal condition holds.

When τ_0 is not identifiable ($\delta_0 = 0$), $\widehat{\tau}$ obtained in the second-step estimation can be any value in $\mathcal{T}$. Note that Lemmas E.4 and E.5 are stated and proved for this case as well. Similar to the proof of Theorem 4.5, by Lemma D.1, in quantile regression models, $M_n = Cs^{1/2}$ for some $C > 0$. Hence all the required conditions in Lemmas E.4 and E.5 are satisfied by the conditions imposed in Theorem 5.2. Then by Lemmas E.4 and E.5, w.p.a.1, for any $\alpha = (\alpha_J, \alpha_{J^c}) \in \mathcal{H}$,

$$\widetilde{S}_n(\bar{\alpha}_J, 0) = \bar{Q}_n(\bar{\alpha}_J) \le \bar{Q}_n(\alpha_J) = \widetilde{S}_n(\alpha_J, 0) \le \widetilde{S}_n(\alpha).$$

Hence $(\bar{\alpha}_J, 0)$ is a local minimizer of $\widetilde{S}_n$, which is also a global minimizer due to the convexity. This implies that w.p.a.1, $\widetilde{\alpha} = (\widetilde{\alpha}_J, \widetilde{\alpha}_{J^c})$ satisfies: $\widetilde{\alpha}_{J^c} = 0$, and $\widetilde{\alpha}_J = \bar{\alpha}_J$, so

$$|\widetilde{\alpha}_J - \alpha_{0J}|_2 = O_P\left(\sqrt{\frac{s\log s}{n}}\right), \quad |\widetilde{\alpha}_J - \alpha_{0J}|_1 = O_P\left(s\sqrt{\frac{\log s}{n}}\right).$$

Also, note that $R(\alpha_0, \widehat{\tau}) = 0$ when $\delta_0 = 0$. Hence by (E.48),

$$R(\widetilde{\alpha}, \widehat{\tau}) \leq 2R(\alpha_0, \widehat{\tau}) + 3\mu_n\bar{D}|\widetilde{\alpha} - \alpha_0|_1 = O_P(\nu_n s\sqrt{\frac{\log s}{n}}).$$

We now prove part (ii) without the minimum signal condition. The proof is very similar to that of Theorem 4.7. Hence we provide the proof briefly. In fact (E.48) still holds by the same argument. But now $R(\alpha_0, \widehat{\tau}) = 0$. Hence for $\Delta = \widetilde{\alpha} - \alpha_0$,

$$R(\widetilde{\alpha}, \widehat{\tau}) + \frac{1}{2}\mu_n|(\widehat{D}\Delta)_{J^c}|_1 \leq \frac{3}{2}\mu_n|\widehat{D}\Delta_J|_1 \leq 2\mu_n\bar{D}\sqrt{s}\|X(\widehat{\tau})^T\Delta\|_2/\sqrt{\phi},$$

where the last inequality follows from Assumption 3. By (E.55), $\left\|X(\widehat{\tau})^T\Delta\right\|_2^2 \leq CR(\widetilde{\alpha}, \widehat{\tau})$, for some $C > 0$. This implies, for some generic constant $C > 0$, $R(\widetilde{\alpha}, \widehat{\tau})^2 \leq \mu_n^2 sCR(\widetilde{\alpha}, \widehat{\tau})$. It follows that

$$R(\widetilde{\alpha}, \widehat{\tau}) \leq \mu_n^2 sC,$$

and

$$|\Delta|_1^2 \leq Cs\|X(\widehat{\tau})\Delta\|_2^2 \leq CsR(\widetilde{\alpha}, \widehat{\tau}) \leq Cs^2\mu_n^2.$$

F   Additional Simulation Results: Different τ_0 and distributions of Q

Tables 9–11 summarize simulation results when the change point τ_0 and the distribution of Q_i vary. We set $\gamma = 0.5$, i.e. median regression, and $n = 200$ for all designs. We consider three different distributions of Q_i: Uniform$[0, 1]$, $N(0, 1)$, and $\chi^2(1)$. The change point parameter τ_0 varies over $0.3, 0.4, \ldots, 0.7$ quantiles of each Q_i distribution. We can confirm the following two results from these simulation studies. First, the performance of $\widehat{\tau}$ measured by the root-mean-squared error depends on the density of Q_i distribution. For instance, it is quite uniform over different τ_0 when Q_i follows Uniform$[0, 1]$. However, when Q_i follows $N(0, 1)$ or $\chi^2(1)$, it performs better when τ_0 is located at a point with higher density of Q_i distribution. Second, the mean squared error of $\widehat{\alpha}$ and the oracle proportion get better when τ_0 smaller. It might be caused by the simulation design, $X_i \cdot 1(Q_i > \tau_0)$, as it will generate less zeros when τ_0 is smaller and help increase the signal from X_i's.

Table 9: Different τ_0 and Q_i dist.: $Q_i \sim Unif[0,1]$

		Excess Risk	$E[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c}$)	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop.
	Oracle 1	0.008	NA	0.016 (NA / NA)	0.282	NA	NA	NA
	Oracle 2	0.017	NA	0.016 (NA / NA)	0.451	0.010	0.951	NA
$\tau_0 = 0.3$	Step 1	0.039	5.557	0.206 (0.182 / 0.024)	0.697	0.011	0.950	0.026
	Step 2	0.040	NA	NA (NA / NA)	0.700	0.011	0.949	NA
	Step 3a	0.038	5.536	0.201 (0.177 / 0.024)	0.687	0.011	0.947	0.026
	Step 3b	0.041	2.042	0.145 (0.134 / 0.011)	0.717	0.012	0.924	0.475
	Oracle 1	0.008	NA	0.014 (NA / NA)	0.287	NA	NA	NA
	Oracle 2	0.017	NA	0.014 (NA / NA)	0.458	0.011	0.956	NA
$\tau_0 = 0.4$	Step 1	0.039	5.590	0.228 (0.201 / 0.027)	0.706	0.011	0.955	0.019
	Step 2	0.037	NA	NA (NA / NA)	0.707	0.011	0.955	NA
	Step 3a	0.034	5.578	0.226 (0.199 / 0.027)	0.695	0.011	0.949	0.018
	Step 3b	0.040	2.203	0.147 (0.131 / 0.017)	0.704	0.011	0.933	0.492
	Oracle 1	0.008	NA	0.012 (NA / NA)	0.287	NA	NA	NA
	Oracle 2	0.018	NA	0.012 (NA / NA)	0.470	0.010	0.951	NA
$\tau_0 = 0.5$	Step 1	0.042	5.698	0.262 (0.230 / 0.032)	0.706	0.011	0.944	0.020
	Step 2	0.042	NA	NA (NA / NA)	0.711	0.011	0.939	NA
	Step 3a	0.041	5.680	0.256 (0.224 / 0.032)	0.696	0.011	0.941	0.020
	Step 3b	0.041	2.343	0.167 (0.142 / 0.025)	0.714	0.011	0.931	0.443
	Oracle 1	0.008	NA	0.013 (NA / NA)	0.295	NA	NA	NA
	Oracle 2	0.017	NA	0.013 (NA / NA)	0.475	0.011	0.947	NA
$\tau_0 = 0.6$	Step 1	0.042	5.869	0.344 (0.303 / 0.041)	0.731	0.013	0.937	0.012
	Step 2	0.042	NA	NA (NA / NA)	0.742	0.013	0.930	NA
	Step 3a	0.039	5.855	0.336 (0.296 / 0.040)	0.730	0.013	0.928	0.012
	Step 3b	0.041	2.467	0.249 (0.204 / 0.046)	0.734	0.012	0.923	0.382
	Oracle 1	0.007	NA	0.012 (NA / NA)	0.280	NA	NA	NA
	Oracle 2	0.018	NA	0.012 (NA / NA)	0.470	0.010	0.949	NA
$\tau_0 = 0.7$	Step 1	0.041	5.978	0.464 (0.407 / 0.057)	0.729	0.012	0.954	0.016
	Step 2	0.042	NA	NA (NA / NA)	0.737	0.012	0.951	NA
	Step 3a	0.041	5.981	0.456 (0.400 / 0.056)	0.718	0.012	0.953	0.020
	Step 3b	0.040	2.549	0.386 (0.303 / 0.083)	0.706	0.012	0.944	0.319

Note: For all designs, $J(\alpha_\gamma) = 2$, $\gamma = 0.5$, and $n = 200$. See the note below Table 1 for other notation.

Table 10: Different τ_0 and Q_i dist.: $Q_i \sim N(0,1)$

		Excess Risk	$E[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c}$)	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop.
$\tau_0 = -0.52$	Oracle 1	0.008	NA	0.017 (NA / NA)	0.294	NA	NA	NA
	Oracle 2	0.018	NA	0.017 (NA / NA)	0.500	0.034	0.949	NA
	Step 1	0.037	5.389	0.191 (0.168 / 0.023)	0.689	0.036	0.953	0.024
	Step 2	0.039	NA	NA (NA / NA)	0.690	0.036	0.943	NA
	Step 3a	0.039	5.382	0.187 (0.164 / 0.023)	0.677	0.036	0.938	0.023
	Step 3b	0.042	2.248	0.132 (0.125 / 0.008)	0.695	0.042	0.918	0.523
$\tau_0 = -0.25$	Oracle 1	0.008	NA	0.014 (NA / NA)	0.292	NA	NA	NA
	Oracle 2	0.018	NA	0.014 (NA / NA)	0.482	0.028	0.954	NA
	Step 1	0.041	5.722	0.231 (0.204 / 0.027)	0.708	0.027	0.950	0.022
	Step 2	0.034	NA	NA (NA / NA)	0.719	0.028	0.945	NA
	Step 3a	0.039	5.724	0.226 (0.199 / 0.027)	0.717	0.028	0.943	0.022
	Step 3b	0.042	2.231	0.145 (0.129 / 0.016)	0.702	0.029	0.938	0.474
$\tau_0 = 0$	Oracle 1	0.008	NA	0.013(NA / NA)	0.291	NA	NA	NA
	Oracle 2	0.016	NA	0.013 (NA / NA)	0.464	0.025	0.968	NA
	Step 1	0.038	5.709	0.275 (0.242 / 0.033)	0.709	0.028	0.953	0.024
	Step 2	0.040	NA	NA (NA / NA)	0.706	0.028	0.957	NA
	Step 3a	0.038	5.682	0.271 (0.238 / 0.033)	0.691	0.028	0.956	0.023
	Step 3b	0.042	2.309	0.184 (0.156 / 0.029)	0.711	0.027	0.948	0.458
$\tau_0 = 0.25$	Oracle 1	0.008	NA	0.012 (NA / NA)	0.292	NA	NA	NA
	Oracle 2	0.017	NA	0.012 (NA / NA)	0.474	0.028	0.958	NA
	Step 1	0.041	5.829	0.359 (0.316 / 0.043)	0.718	0.029	0.959	0.016
	Step 2	0.043	NA	NA (NA / NA)	0.732	0.030	0.949	NA
	Step 3a	0.039	5.841	0.351 (0.308 / 0.042)	0.730	0.030	0.950	0.016
	Step 3b	0.038	2.456	0.269 (0.219 / 0.050)	0.711	0.030	0.941	0.378
$\tau_0 = 0.52$	Oracle 1	0.008	NA	0.012 (NA / NA)	0.286	NA	NA	NA
	Oracle 2	0.017	NA	0.012(NA / NA)	0.466	0.031	0.964	NA
	Step 1	0.043	5.929	0.455 (0.400 / 0.055)	0.759	0.034	0.953	0.012
	Step 2	0.041	NA	NA (NA / NA)	0.748	0.034	0.947	NA
	Step 3a	0.037	5.932	0.445 (0.390 / 0.055)	0.736	0.033	0.945	0.010
	Step 3b	0.042	2.529	0.395 (0.310 / 0.084)	0.750	0.033	0.940	0.300

Note: For all designs, $J(\alpha_\gamma) = 2$, $\gamma = 0.5$, and $n = 200$. Note that $Quant_{0.3}(Q_i) \approx -0.52$, $Quant_{0.4}(Q_i) \approx -0.25$, $Quant_{0.5}(Q_i) = 0$. See the note below Table 1 for other notation.

Table 11: Different τ_0 and Q_i dist.: $Q_i \sim \chi^2(1)$

		Excess Risk	$E[J(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$ ($\widehat{\alpha}_{J_0}/\widehat{\alpha}_{J_0^c}$)	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop.
$\tau_0 = 0.15$	Oracle 1	0.008	NA	0.017 (NA / NA)	0.293	NA	NA	NA
	Oracle 2	0.017	NA	0.017(NA / NA)	0.461	0.012	0.978	NA
	Step 1	0.038	5.523	0.211 (0.187 / 0.025)	0.701	0.012	0.979	0.032
	Step 2	0.034	NA	NA (NA / NA)	0.721	0.011	0.980	NA
	Step 3a	0.036	5.524	0.207 (0.182 / 0.025)	0.697	0.011	0.980	0.029
	Step 3b	0.037	2.023	0.137 (0.126 / 0.010)	0.692	0.012	0.966	0.523
$\tau_0 = 0.27$	Oracle 1	0.008	NA	0.014 (NA / NA)	0.286	NA	NA	NA
	Oracle 2	0.017	NA	0.014 (NA / NA)	0.448	0.015	0.957	NA
	Step 1	0.036	5.562	0.229 (0.202 / 0.027)	0.720	0.016	0.951	0.026
	Step 2	0.036	NA	NA (NA / NA)	0.712	0.016	0.950	NA
	Step 3a	0.038	5.558	0.225 (0.199 / 0.027)	0.694	0.016	0.947	0.028
	Step 3b	0.040	2.206	0.138 (0.124 / 0.014)	0.693	0.015	0.945	0.507
$\tau_0 = 0.45$	Oracle 1	0.008	NA	0.011 (NA / NA)	0.291	NA	NA	NA
	Oracle 2	0.016	NA	0.011 (NA / NA)	0.461	0.022	0.942	NA
	Step 1	0.036	5.810	0.291 (0.256 / 0.035)	0.718	0.022	0.934	0.017
	Step 2	0.038	NA	NA (NA / NA)	0.722	0.021	0.930	NA
	Step 3a	0.041	5.834	0.288 (0.253 / 0.035)	0.706	0.021	0.930	0.019
	Step 3b	0.041	2.353	0.207 (0.171 / 0.036)	0.712	0.021	0.919	0.439
$\tau_0 = 0.71$	Oracle 1	0.009	NA	0.012 (NA / NA)	0.288	NA	NA	NA
	Oracle 2	0.018	NA	0.012 (NA / NA)	0.485	0.030	0.933	NA
	Step 1	0.035	5.883	0.348 (0.307 / 0.042)	0.717	0.031	0.934	0.015
	Step 2	0.042	NA	NA (NA / NA)	0.741	0.032	0.923	NA
	Step 3a	0.038	5.866	0.337 (0.296 / 0.041)	0.726	0.032	0.922	0.014
	Step 3b	0.038	2.386	0.240 (0.197 / 0.044)	0.724	0.032	0.909	0.397
$\tau_0 = 1.07$	Oracle 1	0.008	NA	0.013(NA / NA)	0.291	NA	NA	NA
	Oracle 2	0.017	NA	0.013 (NA / NA)	0.473	0.044	0.936	NA
	Step 1	0.043	5.967	0.459 (0.404 / 0.055)	0.740	0.049	0.923	0.008
	Step 2	0.041	NA	NA (NA / NA)	0.752	0.050	0.922	NA
	Step 3a	0.036	5.932	0.445 (0.390 / 0.054)	0.738	0.050	0.920	0.010
	Step 3b	0.044	2.486	0.381 (0.303 / 0.078)	0.740	0.048	0.918	0.317

Note: For all designs, $J(\alpha_\gamma) = 2$, $\gamma = 0.5$, and $n = 200$. Note that τ_0 values are $0.3, 0.4, \ldots, 0.7$ quantiles of $\chi^2(1)$. See the note below Table 1 for other notation.

G Additional Simulation Results: Sensitivity Analyses

Tables 12–21 summarize the simulation results of sensitivity analysis on tuning parameters. We set $\gamma = 0.5$, i.e. median regression, and $n = 200$ for all designs. We make variation on four constants of tuning parameters: γ^* of $\overline{\Lambda}_{1-\gamma^*}$, c_1 of κ_n and ω_n, c_2 of μ_n, and a of the signal adaptive weight w_j. Recall that they are set to $\gamma^* = 0.1$, $c_1 = 1.1$, $c_2 = \log\log n$, and $a = 3.7$ following the existing literature and some preliminary simulations. We make changes over the range between -15% and $+15\%$ of the suggested values. Since γ^* and c_1 are relevant for all estimation steps, we report the sensitivity analysis results for all steps: Tables 12–15 and Tables 16–19. However, we only report the results of Step 3b for c_2 and a as they affect only the last step: Table 20 and Table 21. These simulation studies confirm that the proposed estimators are robust to some variation in tuning parameters. Both γ^* and c_1 show some tendency that a smaller penalty size (larger γ^* and smaller c_1) improves the prediction error slightly. Table 20 shows quite stable oracle proportions unless c_2 is too small. The constant a for the signal adaptive weight shows quite uniform performance over different values. Figures 3–12 present graphical representation of the sensitivity analyses reported in Tables 12–21.

Table 12: Sensitivity Analysis of γ^*: Step 1

Changes	Excess Risk	$\mathrm{E}[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.039	5.776	0.279	0.246	0.033	0.730	0.011	0.945	0.015
-12%	0.039	5.717	0.280	0.247	0.033	0.727	0.011	0.937	0.019
-9%	0.040	5.720	0.278	0.245	0.033	0.739	0.012	0.931	0.013
-6%	0.039	5.635	0.277	0.244	0.033	0.732	0.012	0.945	0.016
-3%	0.040	5.767	0.282	0.248	0.034	0.733	0.012	0.940	0.017
0%	0.040	5.790	0.279	0.245	0.034	0.723	0.011	0.950	0.015
+3%	0.043	5.782	0.275	0.242	0.033	0.741	0.011	0.944	0.017
+6%	0.038	5.711	0.272	0.239	0.033	0.715	0.011	0.944	0.018
+9%	0.041	5.745	0.278	0.245	0.034	0.697	0.010	0.961	0.018
+12%	0.040	5.730	0.272	0.240	0.032	0.735	0.010	0.957	0.012
+15%	0.042	5.809	0.271	0.240	0.032	0.713	0.011	0.949	0.010

Table 13: Sensitivity Analysis of γ^*: Step 2

Changes	Excess Risk	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$
-15%	0.035	0.726	0.011	0.937
-12%	0.035	0.721	0.011	0.937
-9%	0.036	0.731	0.012	0.933
-6%	0.036	0.727	0.012	0.934
-3%	0.042	0.719	0.012	0.931
0%	0.036	0.729	0.011	0.946
+3%	0.044	0.728	0.012	0.936
+6%	0.038	0.719	0.011	0.936
+9%	0.040	0.703	0.010	0.956
+12%	0.039	0.701	0.010	0.948
+15%	0.040	0.720	0.011	0.946

Table 14: Sensitivity Analysis of γ^*: Step 3a

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.040	5.793	0.277	0.244	0.033	0.722	0.011	0.941	0.014
-12%	0.040	5.750	0.275	0.243	0.032	0.716	0.011	0.938	0.022
-9%	0.041	5.732	0.273	0.240	0.033	0.726	0.012	0.934	0.015
-6%	0.040	5.699	0.272	0.239	0.033	0.724	0.012	0.935	0.017
-3%	0.040	5.795	0.278	0.245	0.034	0.729	0.012	0.933	0.018
0%	0.039	5.910	0.272	0.239	0.033	0.717	0.011	0.947	0.017
+3%	0.036	5.782	0.269	0.237	0.033	0.714	0.012	0.938	0.018
+6%	0.039	5.726	0.267	0.235	0.033	0.699	0.011	0.938	0.018
+9%	0.040	5.747	0.272	0.239	0.033	0.688	0.010	0.959	0.017
+12%	0.038	5.740	0.267	0.236	0.032	0.707	0.010	0.949	0.012
+15%	0.034	5.836	0.267	0.235	0.032	0.703	0.011	0.944	0.009

Table 15: Sensitivity Analysis of γ^*: Step 3b

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.041	2.331	0.185	0.157	0.028	0.726	0.011	0.932	0.429
-12%	0.040	2.309	0.179	0.153	0.025	0.720	0.011	0.928	0.447
-9%	0.042	2.338	0.195	0.165	0.030	0.734	0.012	0.922	0.417
-6%	0.042	2.296	0.189	0.161	0.028	0.732	0.012	0.924	0.449
-3%	0.043	2.311	0.186	0.158	0.028	0.721	0.012	0.924	0.435
0%	0.040	2.330	0.182	0.155	0.027	0.702	0.011	0.929	0.428
+3%	0.041	2.333	0.176	0.150	0.027	0.725	0.012	0.922	0.434
+6%	0.037	2.299	0.173	0.146	0.026	0.689	0.011	0.932	0.467
+9%	0.040	2.325	0.177	0.150	0.027	0.696	0.010	0.944	0.450
+12%	0.036	2.300	0.170	0.144	0.025	0.682	0.011	0.931	0.465
+15%	0.038	2.326	0.180	0.150	0.029	0.686	0.011	0.931	0.455

Table 16: Sensitivity Analysis of c_1: Step 1

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.039	5.811	0.266	0.233	0.033	0.695	0.011	0.948	0.027
-12%	0.034	5.656	0.273	0.239	0.034	0.707	0.012	0.946	0.010
-9%	0.039	5.870	0.273	0.241	0.032	0.699	0.010	0.964	0.014
-6%	0.036	5.787	0.274	0.241	0.033	0.707	0.009	0.965	0.016
-3%	0.043	5.807	0.274	0.242	0.032	0.716	0.011	0.944	0.008
0%	0.040	5.790	0.279	0.245	0.034	0.723	0.011	0.950	0.015
+3%	0.039	5.736	0.281	0.248	0.033	0.730	0.011	0.951	0.016
+6%	0.040	5.727	0.287	0.252	0.035	0.734	0.011	0.945	0.011
+9%	0.042	5.846	0.284	0.251	0.033	0.745	0.011	0.939	0.015
+12%	0.047	5.952	0.309	0.274	0.035	0.753	0.012	0.947	0.012
+15%	0.041	5.828	0.291	0.257	0.033	0.734	0.010	0.961	0.013

Table 17: Sensitivity Analysis of c_1: Step 2

Changes	Excess Risk	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$
-15%	0.038	0.697	0.011	0.943
-12%	0.037	0.713	0.012	0.943
-9%	0.036	0.698	0.010	0.953
-6%	0.034	0.711	0.009	0.954
-3%	0.040	0.723	0.011	0.941
0%	0.036	0.729	0.011	0.946
+3%	0.035	0.724	0.011	0.943
+6%	0.040	0.742	0.011	0.944
+9%	0.038	0.753	0.011	0.931
+12%	0.044	0.753	0.011	0.940
+15%	0.046	0.745	0.009	0.960

Table 18: Sensitivity Analysis of c_1: Step 3a

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.035	5.783	0.262	0.230	0.033	0.676	0.011	0.947	0.026
-12%	0.037	5.691	0.270	0.237	0.034	0.696	0.012	0.946	0.011
-9%	0.040	5.854	0.270	0.238	0.032	0.689	0.010	0.956	0.017
-6%	0.039	5.810	0.271	0.238	0.033	0.705	0.009	0.960	0.015
-3%	0.034	5.832	0.268	0.237	0.031	0.706	0.011	0.943	0.010
0%	0.039	5.910	0.272	0.239	0.033	0.717	0.011	0.947	0.017
+3%	0.040	5.747	0.278	0.244	0.033	0.720	0.011	0.944	0.016
+6%	0.041	5.728	0.281	0.246	0.035	0.728	0.011	0.945	0.014
+9%	0.042	5.789	0.280	0.247	0.033	0.715	0.011	0.935	0.015
+12%	0.038	5.952	0.304	0.269	0.035	0.729	0.011	0.945	0.012
+15%	0.038	5.818	0.285	0.252	0.033	0.735	0.010	0.962	0.016

Table 19: Sensitivity Analysis of c_1: Step 3b

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.035	2.341	0.158	0.129	0.028	0.649	0.011	0.937	0.468
-12%	0.038	2.328	0.190	0.156	0.034	0.678	0.012	0.936	0.415
-9%	0.038	2.346	0.185	0.156	0.029	0.676	0.010	0.946	0.442
-6%	0.037	2.332	0.164	0.138	0.026	0.674	0.010	0.943	0.453
-3%	0.039	2.343	0.187	0.157	0.030	0.694	0.011	0.931	0.447
0%	0.040	2.330	0.182	0.155	0.027	0.702	0.011	0.929	0.428
+3%	0.041	2.331	0.184	0.157	0.027	0.723	0.011	0.935	0.436
+6%	0.047	2.309	0.193	0.164	0.028	0.745	0.012	0.932	0.436
+9%	0.045	2.341	0.170	0.148	0.022	0.732	0.011	0.931	0.434
+12%	0.049	2.385	0.225	0.189	0.035	0.757	0.012	0.925	0.427
+15%	0.045	2.354	0.204	0.175	0.029	0.755	0.010	0.944	0.424

Table 20: Sensitivity Analysis of c_2: Step 3b

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.040	2.455	0.173	0.146	0.027	0.699	0.012	0.928	0.409
-12%	0.040	2.416	0.175	0.148	0.028	0.698	0.011	0.929	0.414
-9%	0.041	2.390	0.175	0.148	0.027	0.702	0.011	0.922	0.431
-6%	0.039	2.373	0.167	0.141	0.025	0.695	0.011	0.931	0.432
-3%	0.041	2.349	0.166	0.142	0.025	0.708	0.011	0.932	0.443
0%	0.040	2.330	0.182	0.155	0.027	0.702	0.011	0.929	0.428
+3%	0.042	2.324	0.192	0.162	0.031	0.716	0.010	0.932	0.464
+6%	0.045	2.308	0.186	0.158	0.028	0.738	0.012	0.917	0.460
+9%	0.045	2.273	0.194	0.164	0.030	0.739	0.011	0.932	0.457
+12%	0.047	2.269	0.192	0.163	0.029	0.757	0.011	0.926	0.438
+15%	0.047	2.266	0.189	0.161	0.028	0.762	0.012	0.927	0.446

Table 21: Sensitivity Analysis of a: Step 3b

Changes	Excess Risk	$E[J_0(\widehat{\alpha})]$	MSE of $\widehat{\alpha}$	MSE of $\widehat{\alpha}_{J_0}$	MSE of $\widehat{\alpha}_{J_0^c}$	Pred. Er.	RMSE of $\widehat{\tau}$	C. Prob. of $\widehat{\tau}$	Oracle Prop
-15%	0.040	2.599	0.183	0.152	0.031	0.696	0.011	0.932	0.435
-12%	0.040	2.343	0.176	0.148	0.028	0.699	0.011	0.935	0.436
-9%	0.041	2.356	0.176	0.149	0.028	0.706	0.011	0.929	0.433
-6%	0.041	2.371	0.175	0.148	0.027	0.713	0.011	0.924	0.421
-3%	0.041	2.342	0.166	0.142	0.025	0.710	0.011	0.928	0.450
0%	0.040	2.330	0.182	0.155	0.027	0.702	0.011	0.929	0.428
+3%	0.042	2.349	0.174	0.147	0.027	0.718	0.011	0.927	0.443
+6%	0.042	2.350	0.173	0.148	0.026	0.720	0.011	0.929	0.445
+9%	0.046	2.340	0.182	0.155	0.027	0.735	0.011	0.923	0.449
+12%	0.046	2.327	0.184	0.156	0.028	0.733	0.011	0.922	0.457
+15%	0.044	2.339	0.185	0.157	0.028	0.724	0.011	0.935	0.453

Figure 3: Sensitivity Analysis of γ^*: Step 1

Figure 4: Sensitivity Analysis of γ^*: Step 2

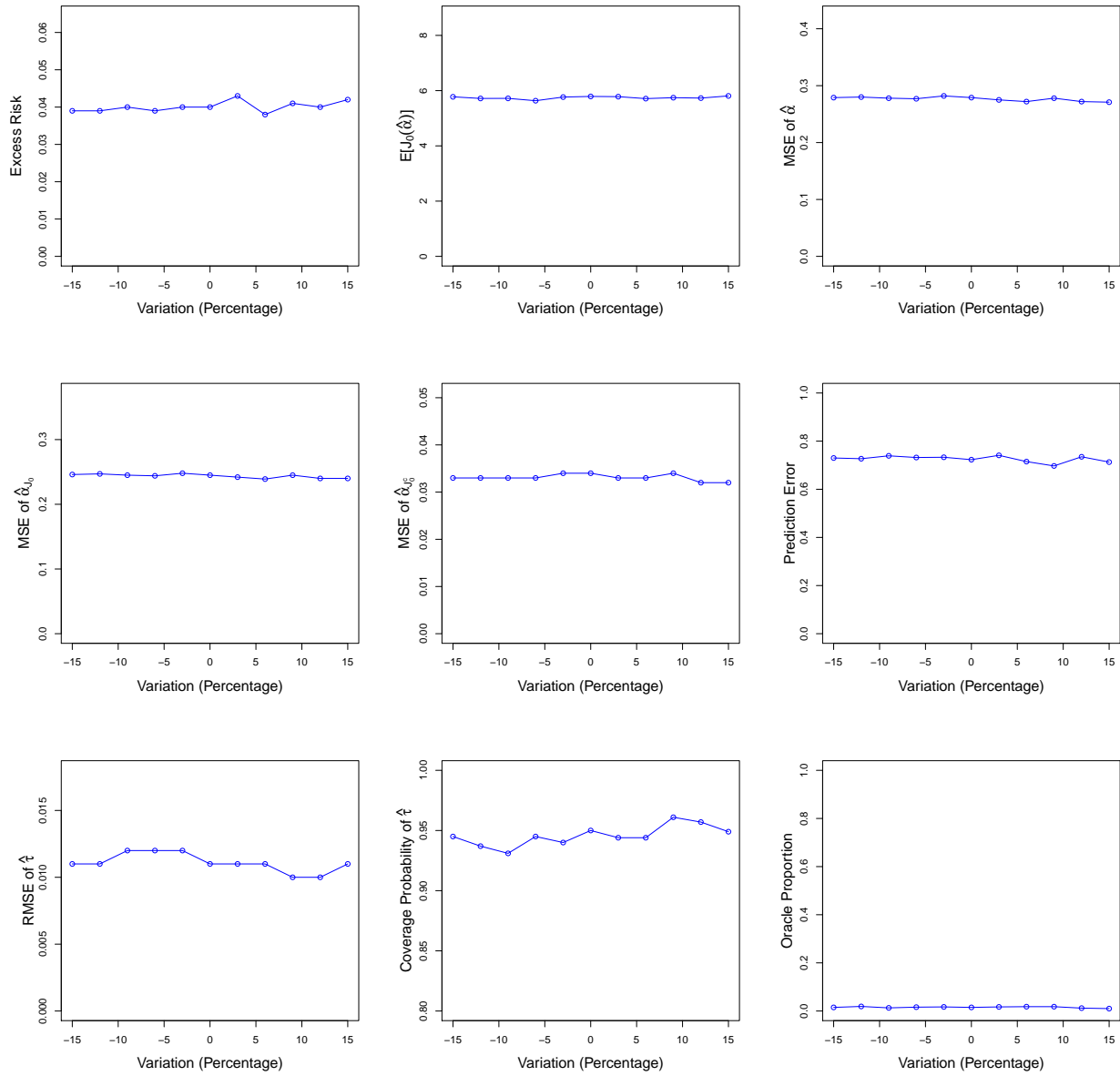Figure 5: Sensitivity Analysis of γ^*: Step 3a

Figure 6: Sensitivity Analysis of γ^*: Step 3b

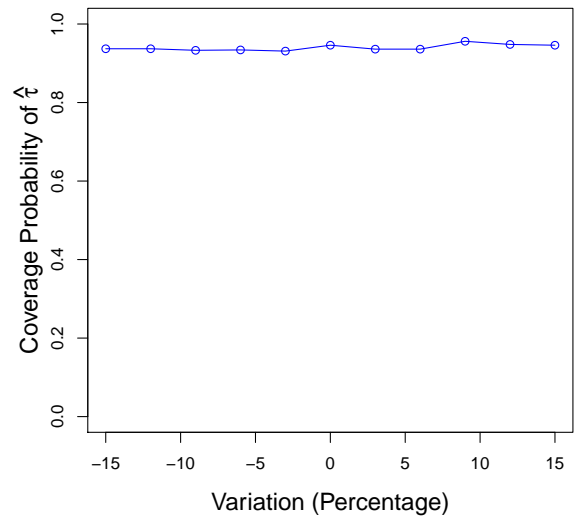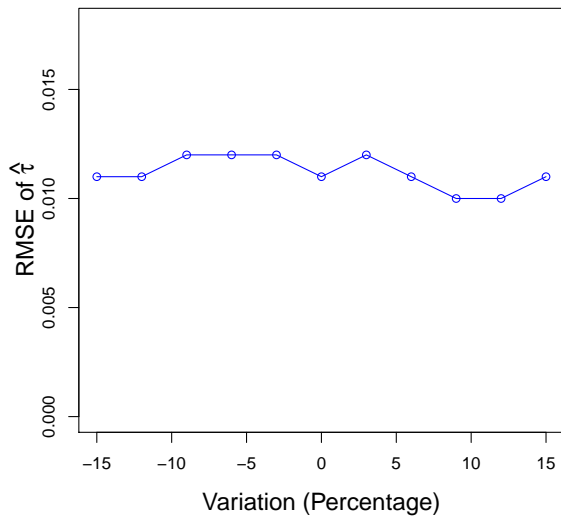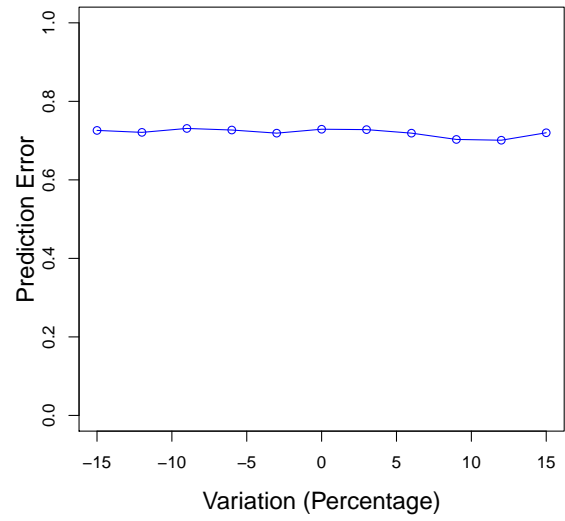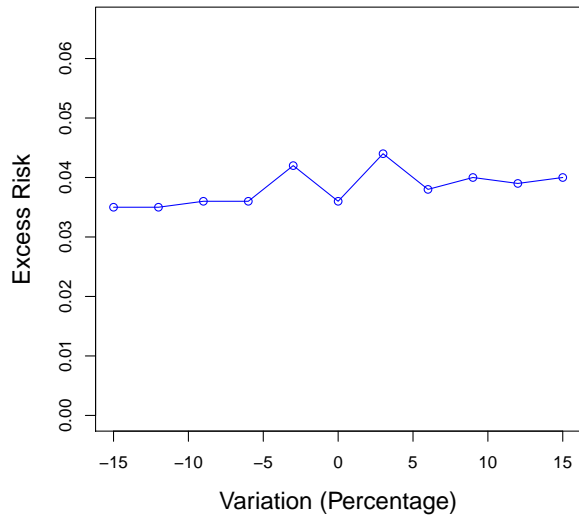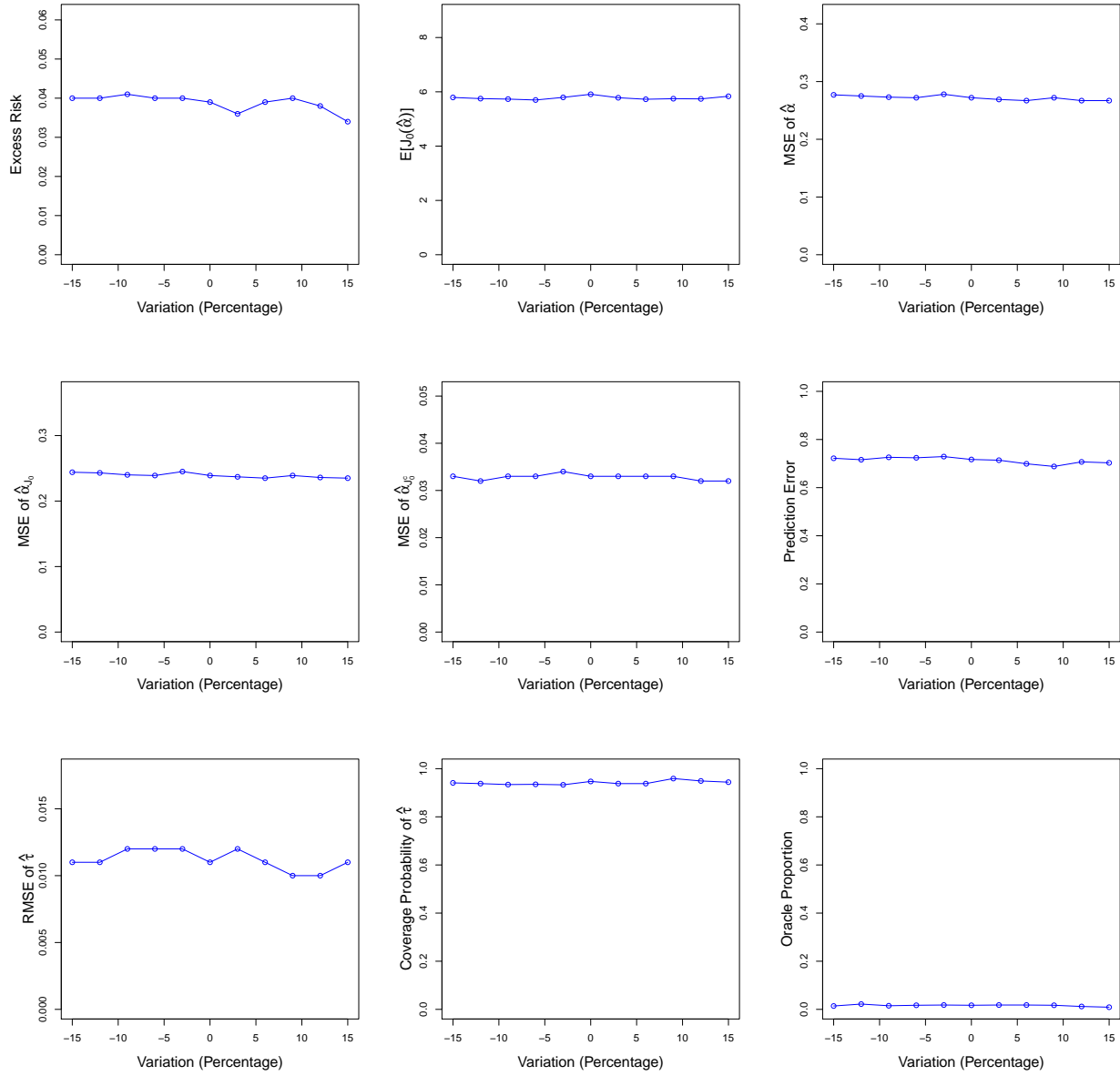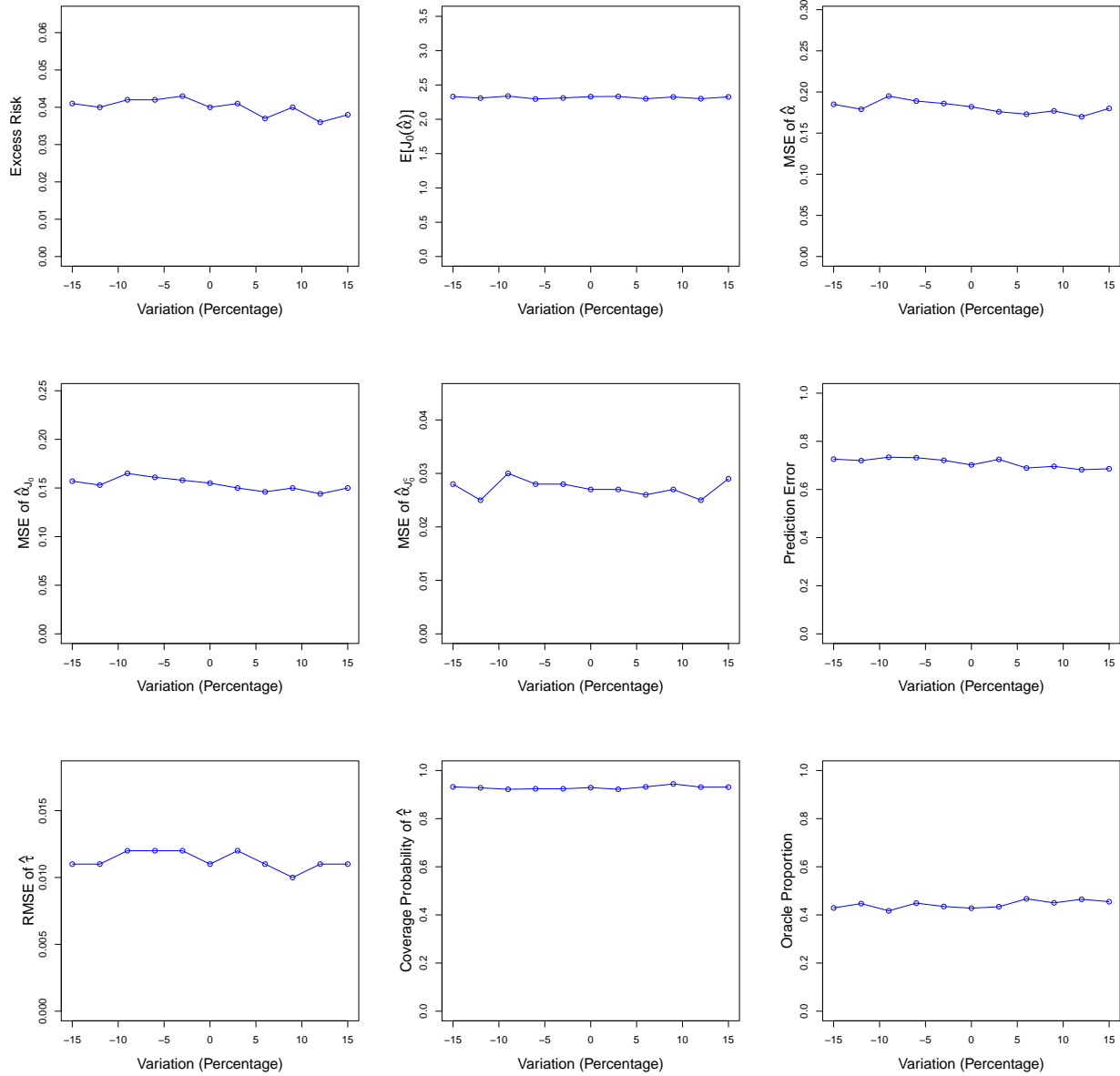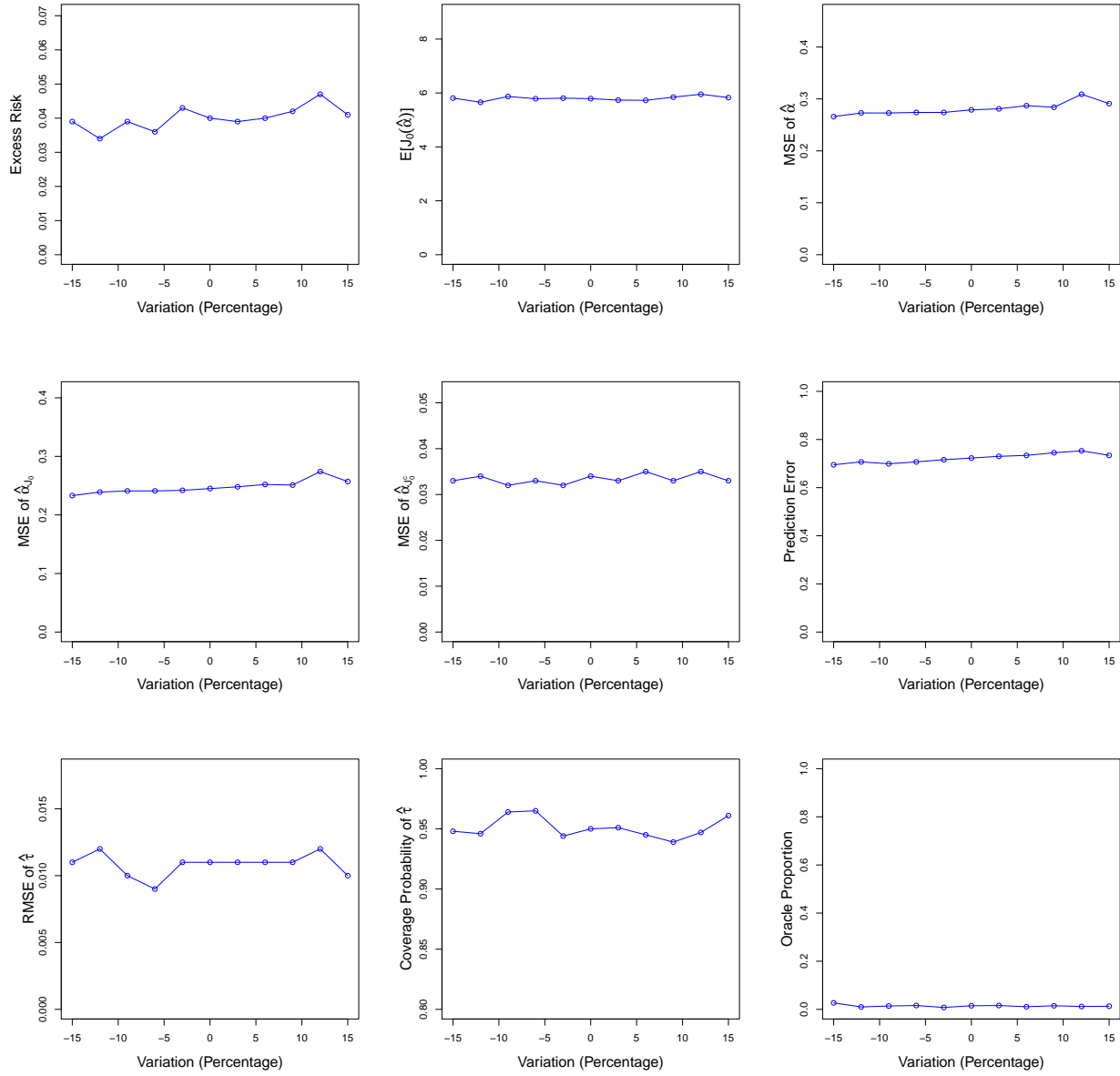Figure 7: Sensitivity Analysis of c_1: Step 1
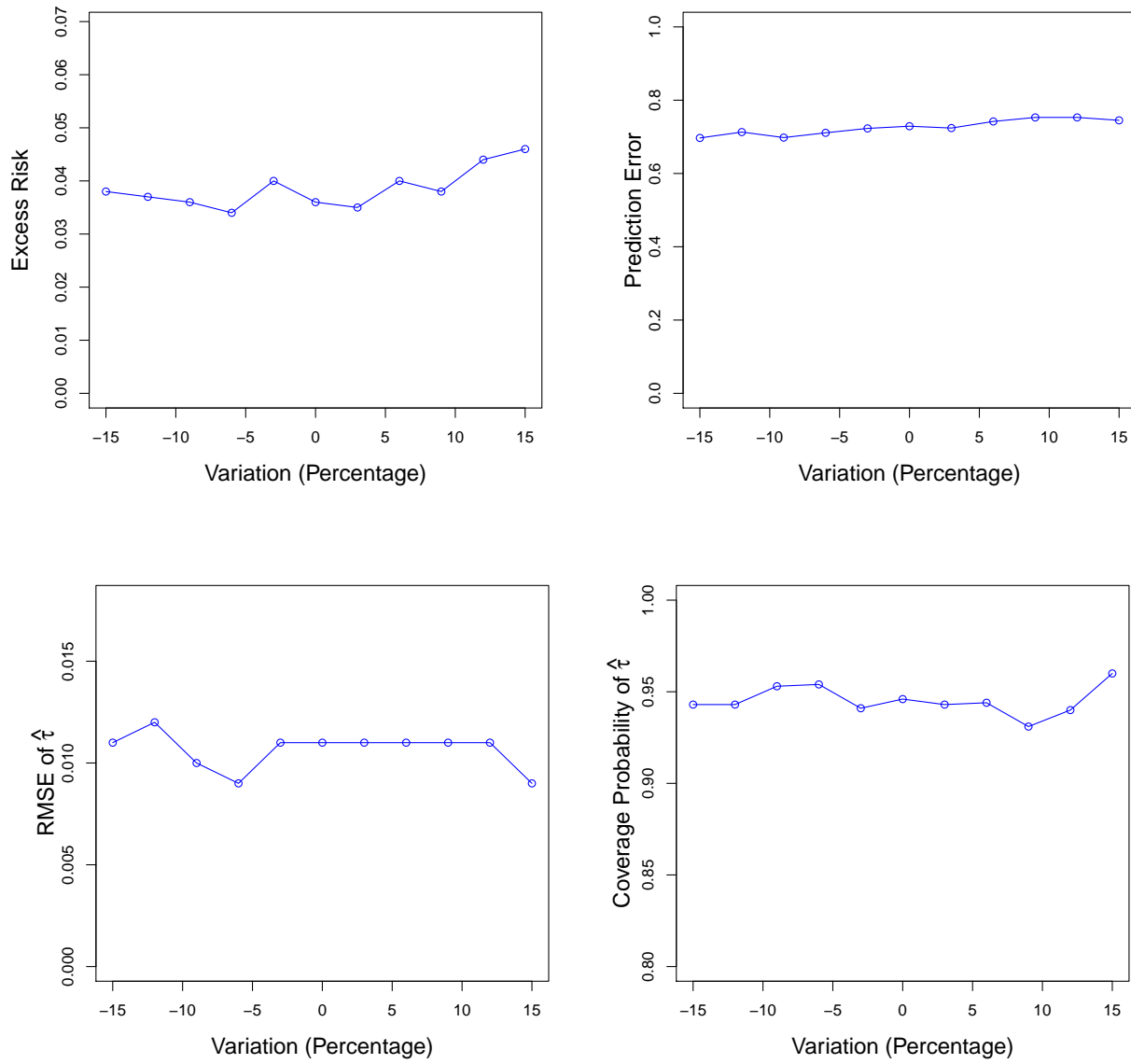
Figure 8: Sensitivity Analysis of c_1: Step 2
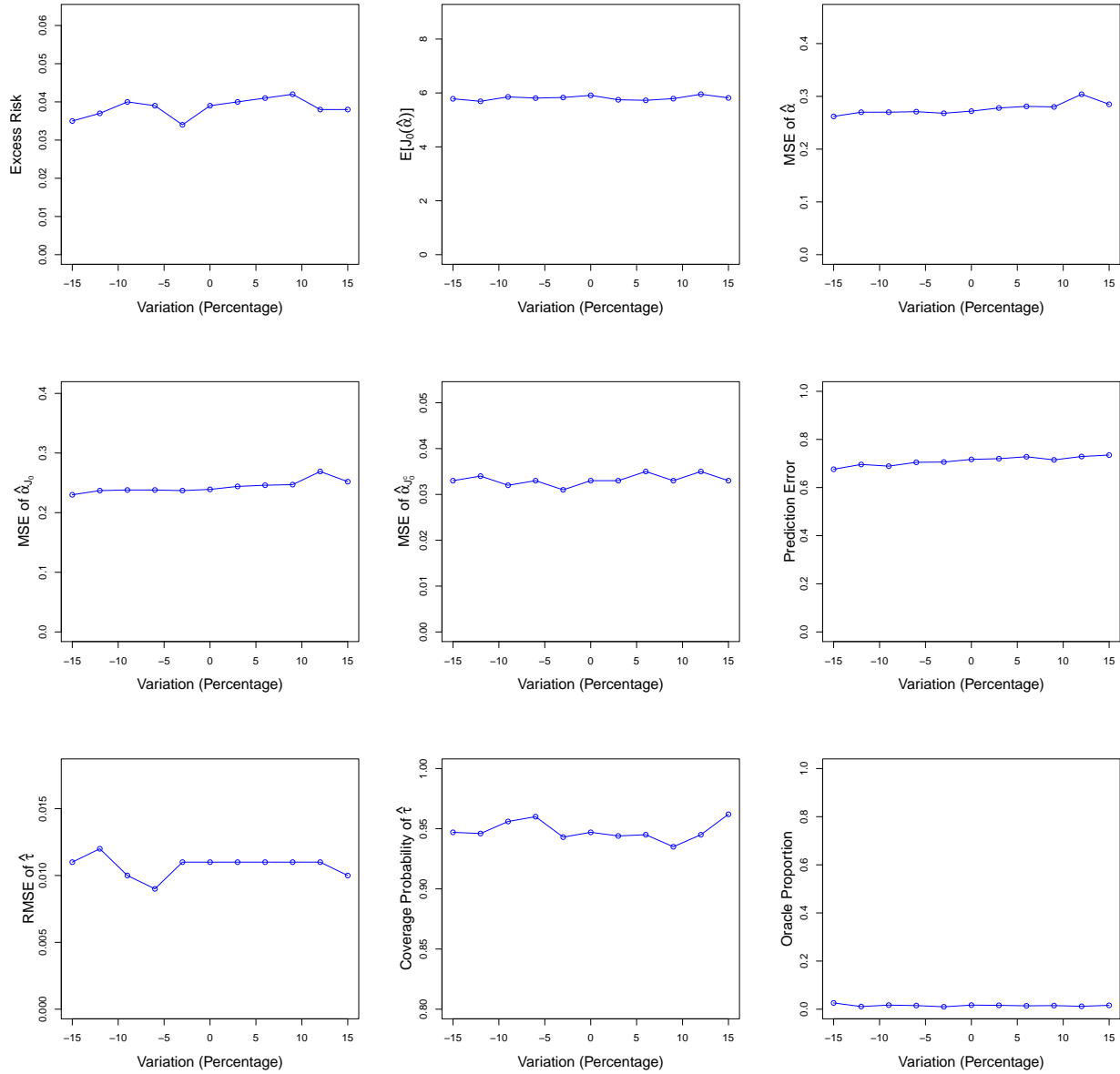
Figure 9: Sensitivity Analysis of c_1: Step 3a
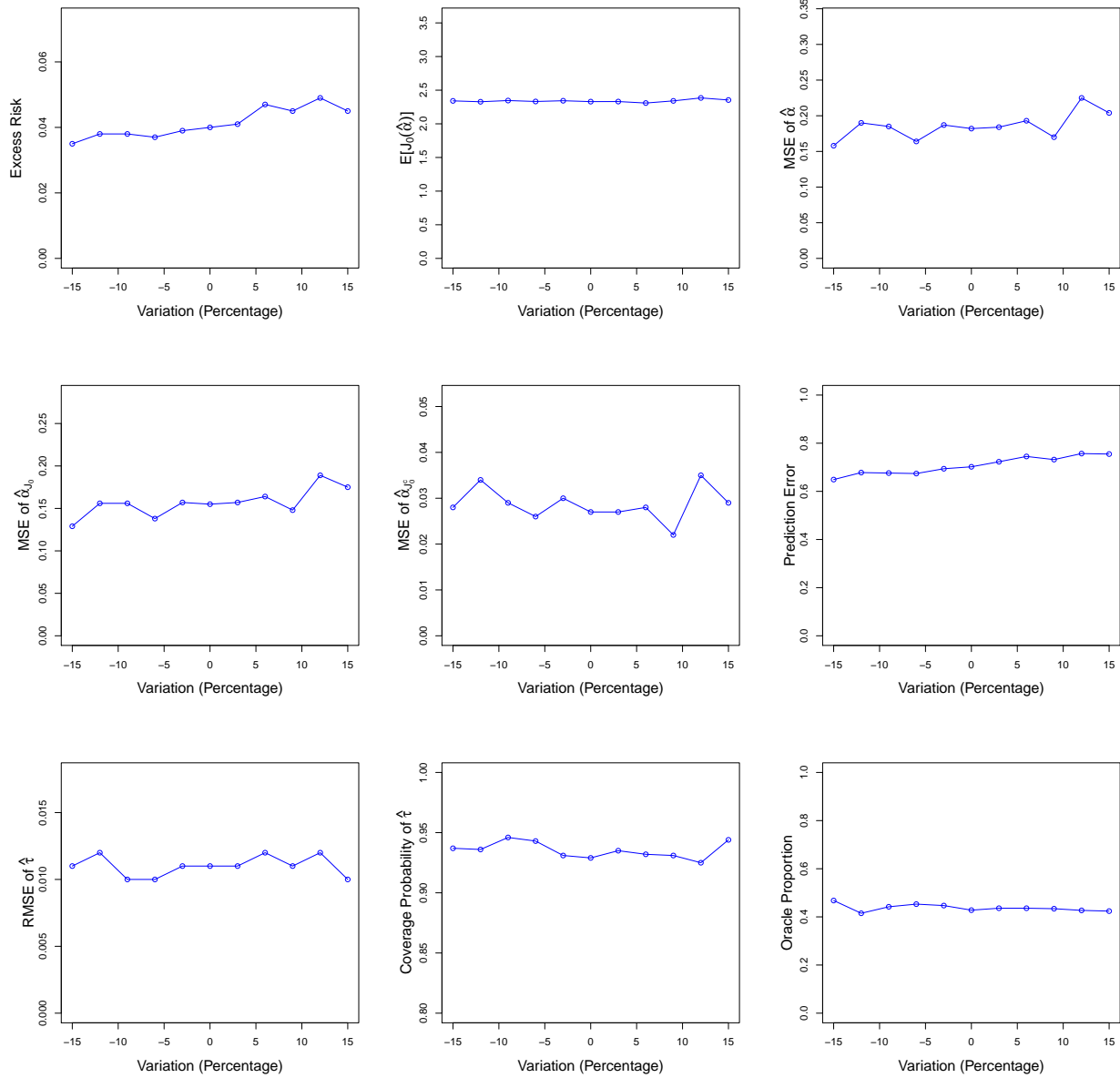
Figure 10: Sensitivity Analysis of c_1: Step 3b
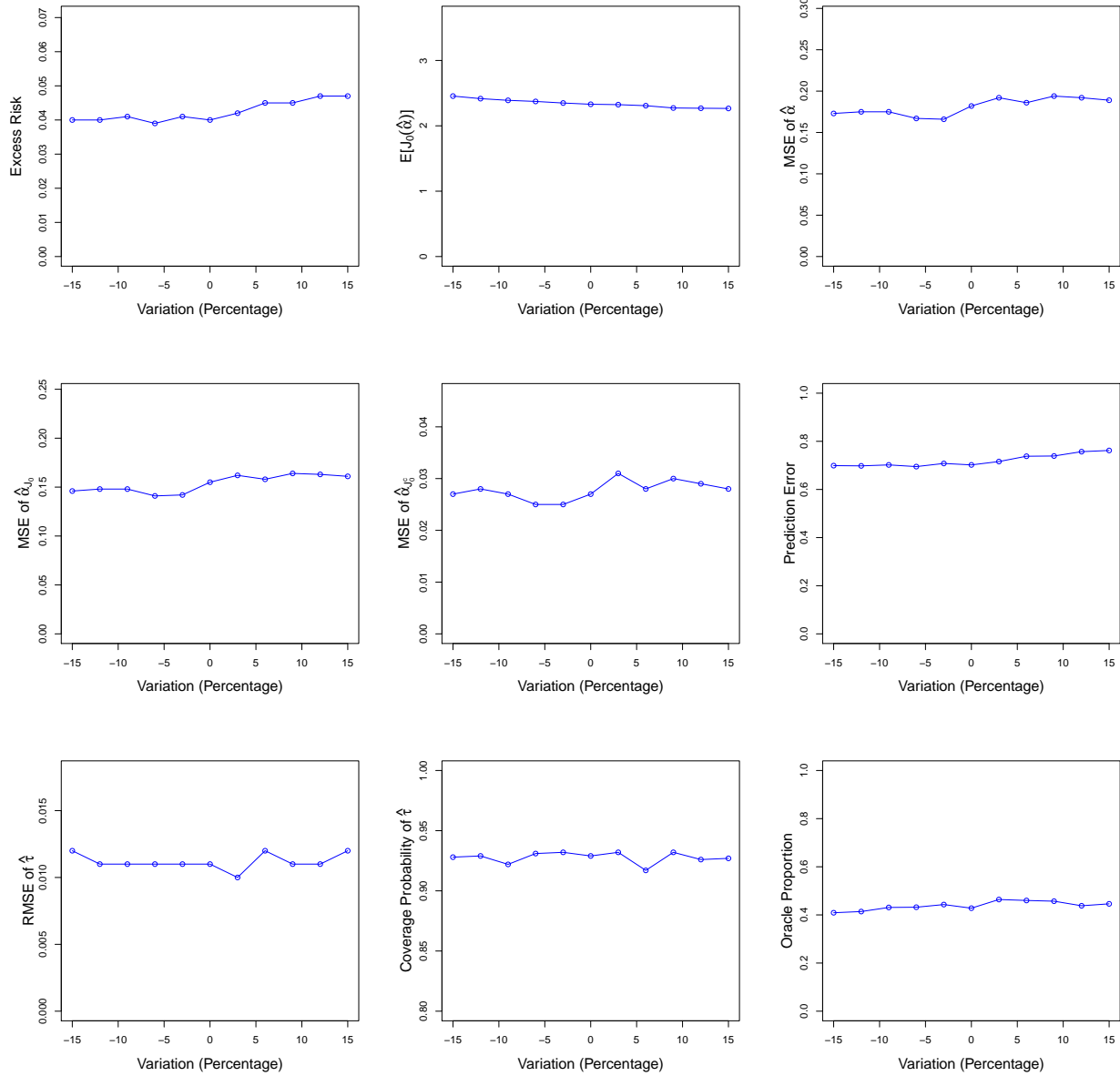
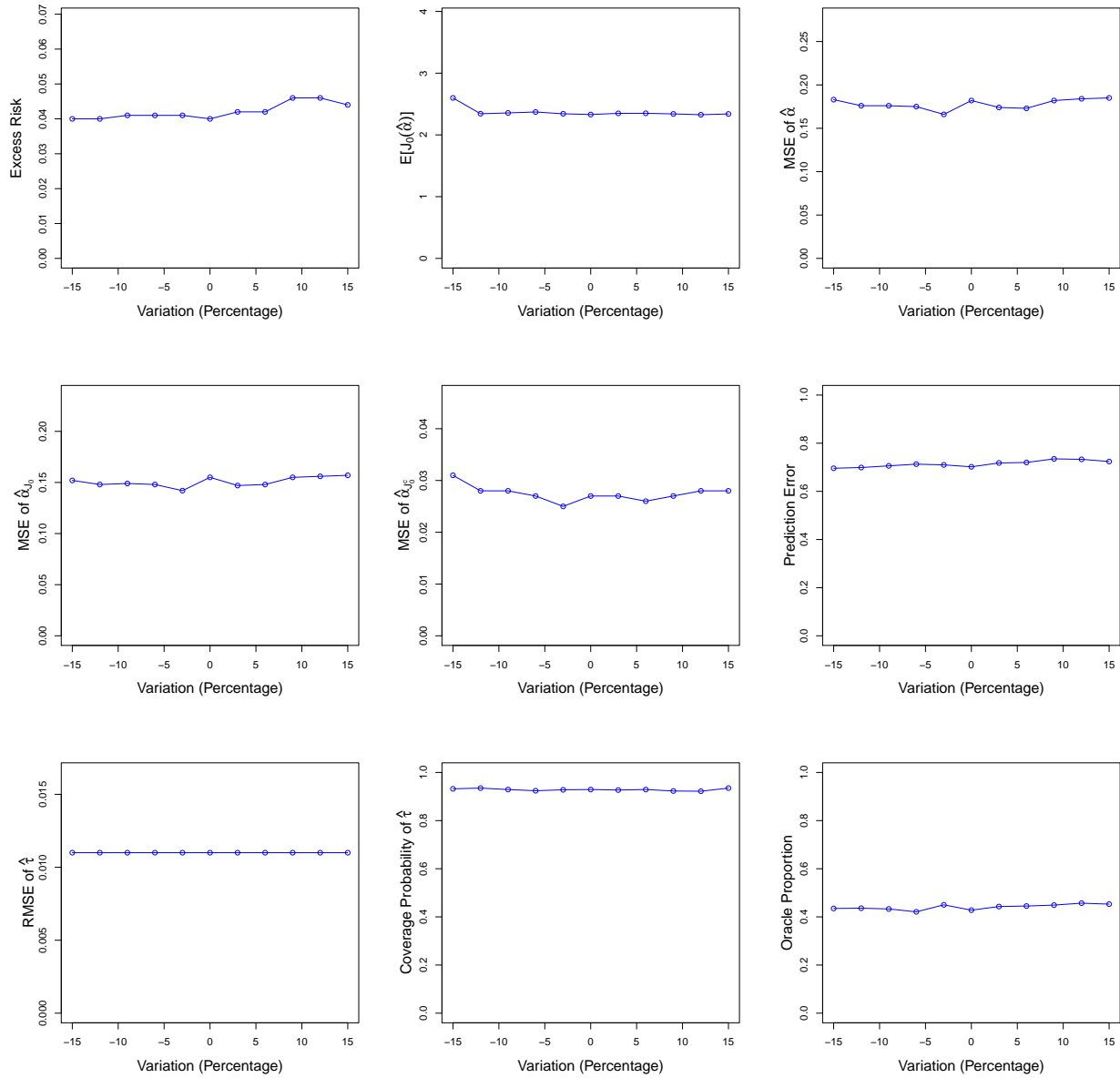Figure 11: Sensitivity Analysis of c_2: Step 3b

Figure 12: Sensitivity Analysis of a: Step 3b

H   Estimating a Change Point in Racial Segregation: Additional Tables

Table 22: Selected Covariates $\gamma = 0.25$

	Selected Regressors
6 control variables	
No Interaction	3 6
Two-way Interaction	1:3 3:5 4:5 5:6
Three-way Interaction	4:5 5:6 1:2:3 1:3:5 1:4:6 2:3:5 3:4:5 3:5:6
Four-way Interaction	4:5 5:6 2:3:5 1:3:4:5 1:4:5:6 3:4:5:6
Five-way Interaction	4:5 5:6 2:3:5 1:4:5:6 1:2:3:4:5 2:3:4:5:6
Six-way Interaction	4:5 5:6 2:3:5 1:4:5:6 1:2:3:4:5 2:3:4:5:6
12 control variables	
No Interaction	3 6
Two-way Interaction	1:5 3:5-sq 5:6 6:5-sq 3-sq:5-sq
Three-way Interaction	5:6 3:5:5-sq 4:1-sq:4-sq 5:1-sq:6-sq 5:3-sq:5-sq 5:4-sq:6-sq 1-sq:3-sq:5-sq 1-sq:5-sq:6-sq
Four-way Interaction	5:6 2:5:6 1:3:1-sq:4-sq 1:5:6:6-sq 2:1-sq:5-sq:6-sq 3:5:2-sq:4-sq 3:5:2-sq:5-sq 3:1-sq:4-sq:6-sq 5:1-sq:2-sq:6-sq 5:1-sq:3-sq:5-sq 5:2-sq:3-sq:5-sq 6:3-sq:4-sq:6-sq
Five-way Interaction	5:6 2:5:6 1:3:4:1-sq:4-sq 1:3:1-sq:4-sq:6-sq 1:5:6:2-sq:6-sq 1:5:3-sq:4-sq:5-sq 1:1-sq:2-sq:5-sq:6-sq 2:5:1-sq:2-sq:6-sq 2:5:2-sq:3-sq:5-sq 3:5:1-sq:3-sq:5-sq 4:6:3-sq:4-sq:6-sq 4:1-sq:2-sq:5-sq:6-sq
Six-way Interaction	2:5:6 1:3:4:1-sq:4-sq 2:5:1-sq:2-sq:6-sq 2:5:2-sq:3-sq:5-sq 1:2:5:6:2-sq:6-sq 1:2:1-sq:2-sq:5-sq:6-sq 1:3:4:5:3-sq:5-sq 1:3:4:6:1-sq:4-sq 1:3:2-sq:4-sq:5-sq:6-sq 1:4:1-sq:3-sq:4-sq:6-sq 2:4:1-sq:2-sq:5-sq:6-sq 3:4:6:3-sq:4-sq:6-sq 3:5:1-sq:2-sq:3-sq:5-sq 4:5:1-sq:2-sq:3-sq:5-sq

Note: Numbers 1 to 6 refer to 6 tract-level control variables: the unemployment rate(1), the log of mean family income(2), the fractions of vacant(3), renter-occupied housing units(4), and single-unit(5), and the fraction of workers who use public transport to travel to work(6). Notation '-sq' stands for the squared variable. The colon (:) denotes interaction between covariates. For example, 1:2 stands for interaction between the unemployment rate and the log of the mean family income.

Table 23: Selected Covariates $\gamma = 0.50$

	Selected Regressors
6 control variables	
No Interaction	1 3 6
Two-way Interaction	1:3 3:5 4:5 5:6
Three-way Interaction	4:5 1:3:5 2:3:5 2:4:6 2:5:6 3:4:5
Four-way Interaction	4:5 5:6 2:3:5 1:3:4:5 1:4:5:6 2:3:4:6
Five-way Interaction	4:5 5:6 2:3:5 1:4:5:6 2:3:4:6 1:2:3:4:5
Six-way Interaction	4:5 5:6 2:3:5 1:4:5:6 2:3:4:6 1:2:3:4:5
12 control variables	
No Interaction	1 3 6 5-sq
Two-way Interaction	1 1:5 3:5-sq 4:5 5:6 6:4-sq 3-sq:5-sq
Three-way Interaction	1 5 1:5 3:5:5-sq 5:6:2-sq 5:1-sq:6-sq 5:4-sq:6-sq 1-sq:3-sq:5-sq 1-sq:5-sq:6-sq 2-sq:3-sq:5-sq 4-sq:5-sq:6-sq
Four-way Interaction	1 5 5:6:2-sq 1:4:5-sq:6-sq 1:5:6:6-sq 1:3-sq:4-sq:5-sq 2:2-sq:3-sq:5-sq 3:5:2-sq:5-sq 3:1-sq:4-sq:6-sq 4:5:6:6-sq 5:1-sq:2-sq:6-sq 5:1-sq:3-sq:5-sq 1-sq:2-sq:5-sq:6-sq
Five-way Interaction	1 5 2:2-sq:3-sq:5-sq 5:1-sq:3-sq:5-sq 1:4:2-sq:5-sq:6-sq 1:5:3-sq:4-sq:5-sq 2:3:5:2-sq:5-sq 2:5:1-sq:2-sq:6-sq 2:5:2-sq:3-sq:5-sq 2:1-sq:2-sq:5-sq:6-sq 3:4:1-sq:4-sq:6-sq 3:5:2-sq:5-sq:6-sq 3:6:1-sq:4-sq:6-sq 4:5:6:2-sq:6-sq 4:1-sq:2-sq:5-sq:6-sq
Six-way Interaction	1 5 2:2-sq:3-sq:5-sq 5:1-sq:3-sq:5-sq 1:5:3-sq:4-sq:5-sq 2:3:5:2-sq:5-sq 2:5:1-sq:2-sq:6-sq 2:5:2-sq:3-sq:5-sq 2:1-sq:2-sq:5-sq:6-sq 1:3:1-sq:3-sq:5-sq:6-sq 1:3:2-sq:4-sq:5-sq:6-sq 2:3:5:2-sq:5-sq:6-sq 2:4:1-sq:2-sq:5-sq:6-sq 3:4:6:1-sq:4-sq:6-sq

Note: Numbers 1 to 6 refer to 6 tract-level control variables: the unemployment rate(1), the log of mean family income(2), the fractions of vacant(3), renter-occupied housing units(4), and single-unit(5), and the fraction of workers who use public transport to travel to work(6). Notation '-sq' stands for the squared variable. The colon (:) denotes interaction between covariates. For example, 1:2 stands for interaction between the unemployment rate and the log of the mean family income.

Table 24: Selected Covariates $\gamma = 0.75$

	Selected Covariates
6 control variables	
No Interaction	3 5
Two-way Interaction	1:3 3:5 4:5 4:6 5:6
Three-way Interaction	1:2 4:5 1:3:5 1:5:6 2:3:5 2:4:5 2:5:6 3:4:5 3:4:6 3:5:6
Four-way Interaction	1:2 4:5 2:3:5 2:5:6 3:5:6 1:3:4:5 1:4:5:6 2:3:4:6
Five-way Interaction	1:2 4:5 2:3:5 2:5:6 3:5:6 1:4:5:6 2:3:4:6 1:2:3:4:5
Six-way Interaction	1:2 4:5 2:3:5 2:5:6 3:5:6 1:4:5:6 2:3:4:6 1:2:3:4:5
12 control variables	
No Interaction	1 3 5-sq
Two-way Interaction	1:2 1:3 1:1-sq 3:5-sq 4:5 5:6 6:6-sq 3-sq:4-sq 3-sq:5-sq 4-sq:6-sq
Three-way Interaction	1 3:5:5-sq 3:1-sq:5-sq 3:2-sq:5-sq 3:5-sq:6-sq 4:3-sq:4-sq 5:6:2-sq
	5:1-sq:6-sq 5:4-sq:6-sq 1-sq:3-sq:5-sq 1-sq:5-sq:6-sq 2-sq:3-sq:5-sq
Four-way Interaction	1 1-sq:3-sq:5-sq 1:3:4:5 1:3:4-sq:5-sq 1:4:5-sq:6-sq 2:3:2-sq:5-sq
	2:5:6:2-sq 2:2-sq:3-sq:5-sq 3:4:2-sq:4-sq 3:5:2-sq:5-sq 3:5:5-sq:6-sq
	3:1-sq:4-sq:6-sq 4:1-sq:5-sq:6-sq 5:1-sq:3-sq:5-sq 1-sq:2-sq:5-sq:6-sq
Five-way Interaction	1 1-sq:3-sq:5-sq 1:3:4:5 1:3:4-sq:5-sq 1:4:5-sq:6-sq 2:3:2-sq:5-sq
	2:5:6:2-sq 2:2-sq:3-sq:5-sq 3:4:2-sq:4-sq 3:5:2-sq:5-sq 3:5:5-sq:6-sq
	3:1-sq:4-sq:6-sq 4:1-sq:5-sq:6-sq 5:1-sq:3-sq:5-sq 1-sq:2-sq:5-sq:6-sq
Six-way Interaction	1 5:6 1-sq:3-sq:5-sq 1:3:4:5 2:3:2-sq:5-sq 2:2-sq:3-sq:5-sq 5:1-sq:3-sq:5-sq 2:3:4:2-sq:4-sq 2:3:5:2-sq:5-sq 2:1-sq:2-sq:5-sq:6-sq 2:3:5:2-sq:5-sq:6-sq 2:4:5:1-sq:3-sq:5-sq 2:4:1-sq:2-sq:5-sq:6-sq 3:4:5:1-sq:2-sq:5-sq 4:6:1-sq:3-sq:4-sq:6-sq

Note: Numbers 1 to 6 refer to 6 tract-level control variables: the unemployment rate(1), the log of mean family income(2), the fractions of vacant(3), renter-occupied housing units(4), and single-unit(5), and the fraction of workers who use public transport to travel to work(6). Notation '-sq' stands for the squared variable. The colon (:) denotes interaction between covariates. For example, 1:2 stands for interaction between the unemployment rate and the log of the mean family income.

Table 25: Full Estimation Results from Mean Regression (Untrimmed Data)

	No. of Reg.	No. of Selected Reg.	$\widehat{\tau}$	CI for τ_0	$\widehat{\delta}$
6 control variables					
No Interaction	26	25	3.25	NA	-21.53
Two-way Interaction	41	34	3.25	NA	-17.39
Three-way Interaction	61	40	3.25	NA	-16.60
Four-way Interaction	76	50	3.25	NA	-15.80
Five-way Interaction	82	49	3.25	NA	-16.12
Six-way Interaction	83	50	3.25	NA	-16.14
12 control variables					
No Interaction	32	29	3.25	NA	-19.94
Two-way Interaction	98	54	3.25	NA	-15.27
Three-way Interaction	318	80	3.25	NA	-15.33
Four-way Interaction	813	103	3.25	NA	-15.16
Five-way Interaction	1605	129	3.25	NA	-15.27
Six-way Interaction	2529	142	3.25	NA	-15.55

Note: The sample size of untrimmed data is $n = 1,813$. The parameter τ_0 is estimated by the grid search on the 591 equi-spaced points over $[1, 60]$. As in the simulation studies, the tuning parameters are set from Step 1 in median regression.

Table 26: Full Estimation Results from Mean Regression (Trimmed Data)

	No. of Reg.	No. of Selected Reg. in Step 3b	$\widehat{\tau}$	CI for τ_0	$\widehat{\delta}$
6 control variables					
No Interaction	26	24	3.35	NA	-6.32
Two-way Interaction	41	32	3.25	NA	-6.00
Three-way Interaction	61	37	3.25	NA	-7.01
Four-way Interaction	76	35	3.25	NA	-6.68
Five-way Interaction	82	41	3.25	NA	-6.59
Six-way Interaction	83	41	3.25	NA	-6.53
12 control variables					
No Interaction	32	30	3.35	NA	-4.27
Two-way Interaction	98	48	3.35	NA	-4.28
Three-way Interaction	318	63	3.25	NA	-5.02
Four-way Interaction	813	90	3.25	NA	-5.18
Five-way Interaction	1605	88	3.25	NA	-5.27
Six-way Interaction	2529	107	3.25	NA	-5.19

Note: The trimmed data drop top and bottom 5% observations based on $\{Y_i\}$ and the sample sizes decreases to $n = 1,626$. The parameter τ_0 is estimated by the grid search on the 591 equi-spaced points over $[1, 60]$. As in the simulation studies, the tuning parameters are set from Step 1 in median regression.

References

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high dimensional sparse models. *Annals of Statistics* **39** 82–130.

BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

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

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data, methods, theory and applications.* Springer, New York.

CALLOT, L., CANER, M., KOCK, A. B. and RIQUELME, J. A. (2016). Sharp threshold detection based on sup-norm error rates in high-dimensional models. *Journal of Business & Economic Statistics* , forthcoming.

CARD, D., MAS, A. and ROTHSTEIN, J. (2008). Tipping and the dynamics of segregation. *Quarterly Journal of Economics* 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.

CHAN, N. H., ING, C.-K., LI, Y. and YAU, C. Y. (2016). Threshold estimation via group orthogonal greedy algorithm. *Journal of Business & Economic Statistics* , forthcoming.

CHAN, N. H., YAU, C. Y. and ZHANG, R.-M. (2014). Group LASSO for structural break time series. *Journal of the American Statistical Association* **109** 590–599.

CHO, H. and FRYZLEWICZ, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 475–507.

CIUPERCA, G. (2013). Quantile regression in high-dimension with breaking. *Journal of Statistical Theory and Applications* **12** 288–305.

ENIKEEVA, F. and HARCHAOUI, Z. (2013). High-dimensional change-point detection with sparse alternatives. *arXiv preprint* http://arxiv.org/abs/1312.1900.

FAN, J., FAN, Y. and BARUT, E. (2014). Adaptive robust variable selection. *Annals of Statistics* **42** 324–351.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

FAN, J. and LV, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory* **57** 5467–5484.

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 495–580.

HANSEN, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64** 413–430.

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

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis* **73** 120–135.

KOENKER, R. (2016). *quantreg: Quantile Regression*. R package version 5.29.
URL https://CRAN.R-project.org/package=quantreg

KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica* 33–50.

KOENKER, R. and MIZERA, I. (2014). Convex optimization in R. *Journal of Statistical Software* **60** 1–23.

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

KOSOROK, M. R. and SONG, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* **35** 957–989.

LEE, S. and SEO, M. H. (2008). Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. *Journal of Econometrics* **144** 492–499.

LEE, S., SEO, M. H. and SHIN, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association* **106** 220–231.

LEE, S., SEO, M. H. and SHIN, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 193–210.

LEONARDI, F. and BÜHLMANN, P. (2016). Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704* http://arxiv.org/abs/1601.03704.

LI, D. and LING, S. (2012). On the least squares estimation of multiple-regime threshold autoregressive models. *Journal of Econometrics* **167** 240–253.

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems 26* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger, eds.). Curran Associates, Inc., 476–484.

Lovász, L. and Vempala, S. (2007). The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms* **30** 307–358.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557.

Pons, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Annals of Statistics* **31** 442–463.

Raskutti, G., Wainwright, M. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57** 6976–6994.

Seijo, E. and Sen, B. (2011a). Change-point in stochastic design regression and the bootstrap. *Annals of Statistics* **39** 1580–1607.

Seijo, E. and Sen, B. (2011b). A continuous mapping theorem for the smallest argmax functional. *Electronic Journal of Statistics* **5** 421–439.

Tong, H. (1990). *Non-linear time series: a dynamical system approach.* Oxford University Press.

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

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes.* Springer, New York.

Wang, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120** 135–151.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Zou, H. and Li, R. (2008). One-step sparse estimations in non concave penalized likelihood models. *Annals of Statistics* **36** 1509–1533.