Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

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

Sokbae Lee, Yuan Liao, Myung Hwan Seo & Youngki Shin

To cite this article: Sokbae Lee, Yuan Liao, Myung Hwan Seo & Youngki Shin (2018) Oracle Estimation of a Change Point in High-Dimensional Quantile Regression, Journal of the American Statistical Association, 113:523, 1184-1194, DOI: 10.1080/01621459.2017.1319840

To link to this article: https://doi.org/10.1080/01621459.2017.1319840

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1184–1194
https://doi.org/10.1080/01621459.2017.1319840

Taylor & Francis
Taylor & Francis Group

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

Sokbae Lee[a], Yuan Liao[b], Myung Hwan Seo[c], and Youngki Shin[d,e]

[a]Department of Economics, Columbia University, New York, NY; Institute for Fiscal Studies, London, United Kingdom; [b]Department of Economics, Rutgers University, New Brunswick, NJ; [c]Department of Economics, Seoul National University, Gwanak-gu, Seoul, Republic of Korea; [d]Economics Discipline Group, University of Technology Sydney, Broadway, NSW, Australia; [e]Department of Economics, McMaster University, Hamilton, Ontario, Canada

ABSTRACT
In this article, 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 homogenous 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 nonsmooth and furthermore the corresponding objective function is nonconvex with respect to the change point. The technique developed in this article 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. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2016
Revised March 2017

KEYWORDS
High-dimensional M-estimation; LASSO; SCAD; Sparsity; Variable selection

1. Introduction

In this article, 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 modeled 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 $\mathbb{P}(U \le 0|X, Q) = \gamma$ 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 heteroscedastic variance or other forms of nonlocation-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 article 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, 2011b; Li and Ling 2012, among many others). Recent articles on change points under high-dimensional setups include Enikeeva and Harchaoui (2013), Chan, Yau, and Zhang (2014), Frick, Munk, and Sieling (2014), Cho and Fryzlewicz (2015), Chan et al. (2017), Callot et al. (2017), and Lee, Seo, and Shin (2016) among others; however, none of these articles consider a change point in high-dimensional quantile regression. The literature on high-dimensional quantile regression includes Belloni and Chernozhukov (2011), Bradic, Fan, and Wang (2011), Wang, Wu, and Li (2012), Wang (2013), and Fan, Fan, and Barut (2014) among others. All the aforementioned articles on quantile regression are under the homogenous sparsity framework (equivalently, assuming that $\delta_0 = 0$ in (1.1)). Ciuperca (2013) considered penalized estimation of a quantile regression model with breaks, but the corresponding analysis is restricted to the case when p is small.

CONTACT Myung Hwan Seo myunghseo@snu.ac.kr Department of Economics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

In this article, 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 article 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 article 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 reestimate 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, Seo, and Shin (2016). However, there are several important differences: first, Lee, Seo, and Shin (2016) considered a high-dimensional mean regression model with a homoscedastic normal error and with deterministic covariates; second, their method consists of one-step least-square estimation with an ℓ_1 penalty; third, they derive nonasymptotic oracle inequalities similar to those in Bickel, Ritov, and Tsybakov (2009) but do not provide any distributional result on the estimator of the change point. Compared to Lee, Seo, and Shin (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-square objective function and is nonsmooth. In addition, we allow for heteroscedastic and nonnormal regression errors and stochastic covariates. These changes coupled with the fact that the quantile regression objective function is nonconvex 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 article 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, Mas, and Rothstein 2008). The empirical question addressed in Card, Mas, and Rothstein (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 5, we use the US Census tract dataset constructed by Card, Mas, and Rothstein (2008) and confirm that the tipping exists in the neighborhoods of Chicago.

The remainder of the article is organized as follows. Section 2 provides an informal description of our estimation methodology. In Section 3.1, we derive the consistency of the estimators in terms of the excess risk. Further asymptotic properties of the proposed estimators are given in Sections 3.2 and 3.3. Section 4 gives a summary of our extensive simulation results. Section 5 illustrates the usefulness of our method by applying it to tipping in the racial segregation and Section 6 concludes. In Appendix A, we provide a set of regularity assumptions to derive asymptotic properties of the proposed estimators in Section 3. Online supplements are comprised of six appendices for all the proofs as well as additional theoretical and numerical results that are left out for the brevity of the article.

Notation. Throughout the article, 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 $\{1, \ldots, 2p\} \setminus J(\alpha_0)$, 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 article.

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 $\rho(t_1, t_2) \equiv (t_1 - t_2)(\gamma - 1\{t_1 - t_2 \leq 0\})$ 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:

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

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 Online 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 ith realization of $X(\tau)$ and jth 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 $D_j(\tau) \equiv \{n^{-1}\sum_{i=1}^{n} X_{ij}(\tau)^2\}^{1/2}$, $j = 1, \dots, 2p$.

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

$$\textit{Step1}: (\breve{\alpha}, \breve{\tau}) = \text{argmin}_{\alpha\in\mathcal{A},\tau\in\mathcal{T}} R_n(\alpha, \tau) + \kappa_n \sum_{j=1}^{2p} D_j(\tau)|\alpha_j|, \quad (2.2)$$

where α_j is the jth element of α. This step produces an initial estimator $(\breve{\alpha}, \breve{\tau})$. The tuning parameter κ_n is required to satisfy

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

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 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. 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; Koenker and Mizera 2014 among others). In other words, the computation cost increases linearly in terms of the number of grid points. In practice, one may choose the grid to be $\{Q_i : i = 1, \dots, n\} \cap \mathcal{T}$.

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:

$$\textit{Step2}: \widehat{\tau} = \underset{\tau\in\mathcal{T}}{\text{argmin}}\, R_n(\breve{\alpha}, \tau), \quad (2.4)$$

where $\breve{\alpha}$ is the first-step estimator of α_0 in (2.2). In Section 3.2, 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 (3.2), define:

Step3a (for prediction):

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

Step3b (for variable selection) :

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

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):

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

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

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

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 3.2 and 3.3, 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 4. 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 reestimate 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 (3.3) indeed holds, a perfect variable selection (variable selection consistency) is possible. The Step 3b estimator achieves the variable selection consistency. In contrast, Step 3a does not use signal-adaptive weights. To achieve the variable selection consistency, it has to rely on much stronger conditions on the design matrix (i.e., the *irrepresentable condition* by 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 6). 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.
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 5), which is useful for predictions.

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 by Belloni and Chernozhukov (2011). Define

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

where U_i is simulated from the iid 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 online Appendix F, 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. In practice, it is important to have a sufficient number of observations lying outside $\mathcal{T}$.

Based on the suggestions by 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; 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. Asymptotic Properties

Throughout the article, 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. In Appendix A, we list a set of assumptions that are needed to derive these properties.

3.1. 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

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

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 following theorem is concerned about the convergence of $R(\breve{\alpha}, \breve{\tau})$ with the first-step estimator.

Theorem 1 (Risk Consistency). Let Assumption A.1 hold. Suppose that the tuning parameter κ_n satisfies (2.3). Then, $R(\breve{\alpha}, \breve{\tau}) = O_P(\kappa_n s)$.

Note that Theorem 1 holds regardless of the identifiability of τ_0 (i.e., 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 chap. 6.6 of Bühlmann and van de Geer 2011) since the objective function $\rho(Y, X(\tau)^T\alpha)$ is nonconvex in τ, due to the unknown change point.

Remark 5. 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 nonidentifiable threshold parameter τ_0. In fact, an inspection of the proof of Theorem 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

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

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

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

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

We first establish the consistency of $\breve{\tau}$ for τ_0.

Theorem 2 (Consistency of $\breve{\tau}$). Let Assumptions A.1, A.2, A.5, and A.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 3 (Rates of Convergence When $\delta_0 \neq 0$). Suppose that $\kappa_n s^2 \log p = o(1)$. Then under Assumptions A.1–A.6, we have

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

In Theorem 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 3 is due to additional assumptions (in particular, compatibility conditions in Assumption A.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, Ritov, and Tsybakov 2009; Raskutti, Wainwright, and Yu 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 article.

Theorem 4 (Oracle Estimation of τ_0). Let Assumptions A.1–A.6 hold. Furthermore, suppose that $\kappa_n s^2 \log p = o(1)$. Then, we have that

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

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

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

where N_1 and N_2 are Poisson processes with the same jump rate $f_Q(\tau_0)$, 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 $\dot{\rho}(U_i - X_i^T\delta_0) - \dot{\rho}(U_i)$ and $\dot{\rho}(U_i + X_i^T\delta_0) - \dot{\rho}(U_i)$ given $Q_i = \tau_0$, where $\dot{\rho}(t) \equiv t(\gamma - 1\{t \leq 0\})$ and $U_i \equiv Y_i - X_i^T\beta_0 - X_i^T\delta_0 1\{Q_i > \tau_0\}$ 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 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 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 6. Li and Ling (2012) proposed a numerical approach for constructing a confidence interval by simulating a compound Poisson process in the context of least-square estimation. We adopt their approach to simulate the compound Poisson process for quantile regression. See Online Appendix B 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 5 (Improved Rates of Convergence When $\delta_0 \neq 0$). Suppose that $\kappa_n s^2 \log p = o(1)$. Then under Assumptions A.1–A.6,

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

Theorem 5 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, for example, Raskutti, Wainwright, and Yu (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 jth 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 6 (Variable Selection When $\delta_0 \neq 0$). Suppose that $\kappa_n s^2 \log p = o(1)$, $s^4 \log s = o(n)$, and

$$\omega_n + s\sqrt{\frac{\log s}{n}} \ll \mu_n \ll \min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|. \tag{3.2}$$

Then under Assumptions A.1–A.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 (3.2) 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 $\widehat{\alpha}_J^* \equiv \text{argmin}_{\alpha_J} R_n^*(\alpha_J, \tau_0)$. 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 7 (Change-Point-Oracle Properties). Suppose that all the conditions imposed in Theorem 6 are satisfied. Furthermore, assume that $\frac{\partial}{\partial \alpha} E[\rho(Y, X^T\alpha)|Q = t]$ 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 $\widetilde{\alpha}_J$, that is, $\mathbf{L}\widetilde{\alpha}_J$, 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 7.

Remark 7. 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 6.

Theorem 8 (Satisfactory Rates Without Minimum Signal Condition). Assume that Assumptions A.1–A.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, $R(\widetilde{\alpha}, \widehat{\tau}) = O_P(\mu_n^2 s)$.

3.3. 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 nonidentifiable, 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 9 (Rates of Convergence When $\delta_0 = 0$). Suppose that $\kappa_n s = o(1)$. Then under Assumptions A.1–A.4, we have that

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

The results obtained in Theorem 9 combined with those obtained in Theorem 3 imply that the first-step estimator 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 5 under Assumptions A.1–A.4. We omit the details for brevity.

We now give a result that is similar to Theorem 6 and Theorem 8.

Theorem 10 (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 A.1–A.4 hold. We have:

(i) If the minimum signal condition holds:

$$\mu_n = o\left(\min_{j \in J(\alpha_0)} |\alpha_{0J}^{(j)}|\right), \tag{3.3}$$

then

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

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

$$P(\widetilde{\beta}_{J^c} = 0) \to 1, \quad P(\widetilde{\delta} = 0) \to 1, \quad \text{and}$$

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

(ii) Without the minimum signal condition (3.3), we have

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

Theorem 10 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 7 can be established straightforwardly as well; however, their details are omitted for brevity.

4. Summary of Monte Carlo Experiments

We have carried out extensive Monte Carlo experiments to examine the finite sample performance of our proposed estimators. To save space, we provide a summary of simulation results and show full results in online Appendix F.

1. The proposed estimator (Step 3b) selected different nonzero coefficients at different quantile levels. The mean regression estimator in Lee, Seo, and Shin (2016) cannot detect these heterogenous models.
2. The coverage probabilities of the confidence interval for τ_0 were good and the root mean squared error of $\widehat{\tau}$ decreased quickly. The latter confirms the superconsistency result of $\widehat{\tau}$.
3. The median regression estimator showed better performances than the mean regression estimator for heteroscedastic designs and for the fat-tail error distributions.
4. The performance of our proposed estimators was satisfactory when $\delta_0 = 0$.
5. When the model contains low *minimal* signals in δ, the Step 3b estimator performed worse than the step 3a estimator.
6. The main qualitative results were not sensitive to different simulation designs on τ_0 and Q_i as well as to some variation on tuning parameter values.

Overall, the simulation results confirm the asymptotic theory developed in the previous sections and show the advantage of quantile regression models over the existing mean regression models with a change point.

5. 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

Table 1. 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

NOTES: The sample size is $n = 1813$. 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 reestimation after Step 3b: that is, τ is estimated again using $(\widetilde{U}_i, \widetilde{\alpha})$ from Step 3b.

constructed by Card, Mas, and Rothstein (2008). They showed 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, Seo, and Shin (2011) developed a test for the existence of threshold effects and applied 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 by Card, Mas, and Rothstein (2008) and Lee, Seo, and Shin (2011) to choose the following median regression with a constant shift due to the tipping effect:

$$Y_i = g_0(Q_i) + \delta_0 1\{Q_i > \tau_0\} + X_i'\beta_0 + U_i, \tag{5.1}$$

where for census tract i, the dependent variable Y_i is the 10 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 jth 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 2529. For example, the number of regressors in the largest model consists of $\#(\text{b-spline basis}) + \#(\text{indicator function}) + \#(\text{interactions up to six-way out of 12}) = 19 + 1 + \sum_{k=1}^{6}\binom{12}{k} = 2529$. This number is much larger than the sample size ($n = 1813$).

Table 1 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 by Card, Mas, and Rothstein (2008). In this empirical example, we report the estimates of τ_0 and the confidence intervals updated after Step 3b (i.e., τ is reestimated 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 G 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

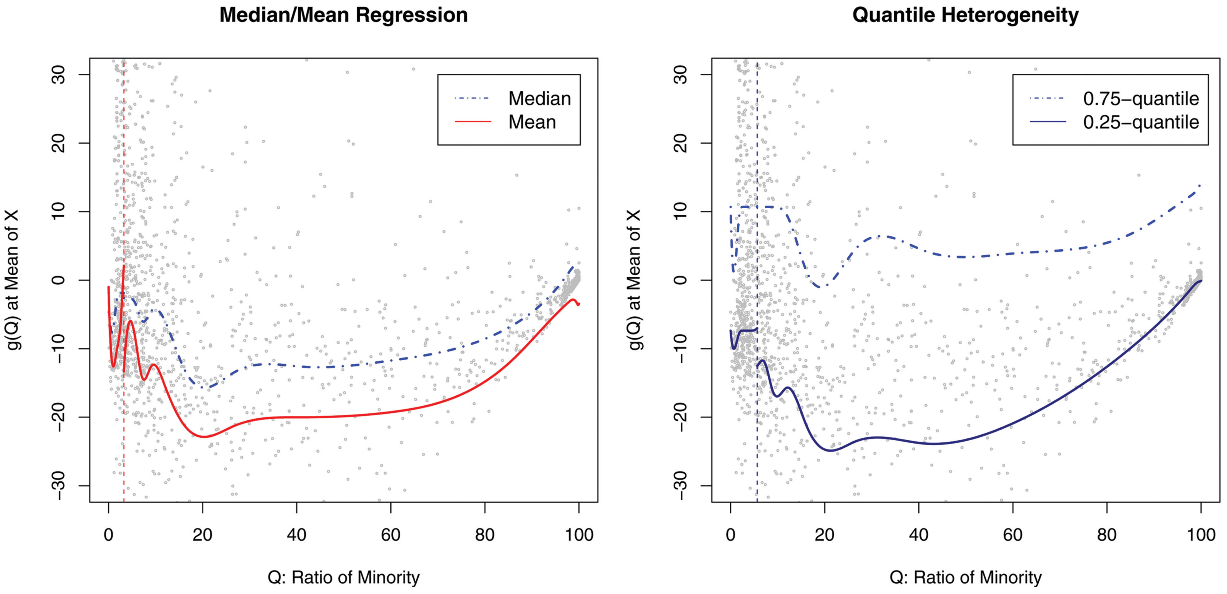

Figure 1. Estimation results: 12 control variables and six-way interaction.

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, that is, $\widehat{\delta} = 0$.

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 and the right panel shows 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. 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.

6. Conclusions

In this article, 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.

One of the important remaining questions is how to extend our approach to a high-dimensional quantile regression model with multiple change points. A computationally attractive approach is to use the binary segmentation algorithm (see, e.g., Fryzlewicz 2014; Cho and Fryzlewicz 2015 among others). In a recent working article, Leonardi and Bühlmann (2016) considered a high-dimensional *mean* regression model with multiple change points whose number may grow as the sample size increases. They have proposed a binary segmentation algorithm to choose the number and locations 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.

Appendix A: Assumptions for Asymptotic Properties

In this section, we list a set of assumptions that will be useful to derive asymptotic properties of the proposed estimators. The first two assumptions are standard.

Assumption A.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 $\mathbb{E}|X_{ij}|^m \leq \frac{m!}{2} K_1^{m-2}$, where X_{ij} denotes the jth element of X_i.

(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 $\tau_0 \in \mathcal{T} \equiv [\underline{\tau}, \overline{\tau}]$. Furthermore, 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) $\mathbb{E}[(X^T\delta_0)^2 | Q = \tau] \leq M_2 |\delta_0|_2^2$ for all $\tau \in \mathcal{T}$ and for some constant M_2 satisfying $0 < M_2 < \infty$.

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.

Assumption A.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.

A.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, chap. 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 $\{1, \ldots, 2p\} \setminus J(\alpha_0)$, respectively.

Assumption A.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 $|\alpha_{J^c}|_1 \leq 5|\alpha_J|_1$,

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

(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 $|\alpha_{J^c}|_1 \leq 4|\alpha_J|_1$,

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

Assumption A.3 requires that the compatibility condition holds uniformly in τ over a neighborhood 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 A.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.

A.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).

We define the "prediction balls" with radius r and corresponding centers as follows:

$$\begin{aligned}\mathcal{B}(\beta_0, r) &= \{\beta \in \mathcal{B} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau_0\}] \le r^2\},\\ \mathcal{G}(\theta_0, r) &= \{\theta \in \mathcal{G} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\theta - \theta_0))^2 1\{Q > \tau_0\}] \le r^2\},\end{aligned} \tag{A.3}$$

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

$$\begin{aligned}\tilde{\mathcal{B}}(\beta_0, r, \tau) &= \{\beta \in \mathcal{B} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\beta - \beta_0))^2 1\{Q \le \tau\}] \le r^2\},\\ \tilde{\mathcal{G}}(\beta_0, r, \tau) &= \{\theta \in \mathcal{G} \subset \mathbb{R}^p : \mathbb{E}[(X^T(\theta - \beta_0))^2 1\{Q > \tau\}] \le r^2\}.\end{aligned} \tag{A.4}$$

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

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

$$\inf_{\beta \in \mathcal{B}(\beta_0, r^*_{QR}), \beta \neq \beta_0} \frac{\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \le \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \le \tau_0\}]} \ge r^*_{QR} \frac{2C_1}{3C_2} > 0 \tag{A.5}$$

and that

$$\inf_{\theta \in \mathcal{G}(\theta_0, r^*_{QR}), \theta \neq \theta_0} \frac{\mathbb{E}[|X^T(\theta - \theta_0)|^2 1\{Q > \tau_0\}]^{3/2}}{\mathbb{E}[|X^T(\theta - \theta_0)|^3 1\{Q > \tau_0\}]} \ge r^*_{QR} \frac{2C_1}{3C_2} > 0. \tag{A.6}$$

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

$$\inf_{\tau \in \mathcal{T}} \inf_{\beta \in \tilde{\mathcal{B}}(\beta_0, r^*_{QR}, \tau), \beta \neq \beta_0} \frac{\mathbb{E}[|X^T(\beta - \beta_0)|^2 1\{Q \le \tau\}]^{3/2}}{\mathbb{E}[|X^T(\beta - \beta_0)|^3 1\{Q \le \tau\}]} \ge r^*_{QR} \frac{2C_1}{3C_2} > 0 \tag{A.7}$$

and that

$$\inf_{\tau \in \mathcal{T}} \inf_{\theta \in \tilde{\mathcal{G}}(\beta_0, r^*_{QR}, \tau), \beta \neq \beta_0} \frac{\mathbb{E}[|X^T(\theta - \theta_0)|^2 1\{Q > \tau\}]^{3/2}}{\mathbb{E}[|X^T(\theta - \theta_0)|^3 1\{Q > \tau\}]} \ge r^*_{QR} \frac{2C_1}{3C_2} > 0. \tag{A.8}$$

Remark A.1. As pointed out by Belloni and Chernozhukov (2011), if $X^T c$ follows a logconcave 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 $\alpha \in \mathcal{A}$,

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

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

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

Assumption A.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 $M_3^{-1} \le \mathbb{E}[(X^T\delta_0)^2 | Q = \tau] \le M_3$ for all $\tau \in \mathcal{T}_0$.

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.

Assumption A.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 $\theta \in \mathcal{G}(\theta_0, r)$,

$$\begin{aligned}\mathbb{E}[|X^T(\beta - \beta_0)| 1\{Q > \tau_0\}] &\le \widetilde{C}\, \mathbb{E}[|X^T(\beta - \beta_0)| 1\{Q \le \tau_0\}],\\ \mathbb{E}[|X^T(\theta - \theta_0)| 1\{Q \le \tau_0\}] &\le \widetilde{C}\, \mathbb{E}[|X^T(\theta - \theta_0)| 1\{Q > \tau_0\}].\end{aligned}$$

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

$$\begin{aligned}&\mathbb{E}\left[(X^T[(\theta - \beta) - (\theta_0 - \beta_0)])^2 \middle| Q = \tau\right] \le M,\\ &\mathbb{E}[|X^T(\beta - \beta_0)| \,|\, Q = \tau] \le M,\\ &\mathbb{E}[|X^T(\theta - \theta_0)| \,|\, Q = \tau] \le M,\\ &\sup_{\tau \in \mathcal{T}_0 : \tau > \tau_0} \mathbb{E}\left[|X^T(\beta - \beta_0)| \frac{1\{\tau_0 < Q \le \tau\}}{(\tau - \tau_0)}\right]\\ &\quad \le M\mathbb{E}[|X^T(\beta - \beta_0)| 1\{Q \le \tau_0\}],\\ &\sup_{\tau \in \mathcal{T}_0 : \tau < \tau_0} \mathbb{E}\left[|X^T(\theta - \theta_0)| \frac{1\{\tau < Q \le \tau_0\}}{(\tau_0 - \tau)}\right]\\ &\quad \le M\mathbb{E}[|X^T(\theta - \theta_0)| 1\{Q > \tau_0\}],\end{aligned}$$

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

Remark A.2. Condition (i) requires that Q have nonnegligible 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 $\mathbb{E}[XX^T 1\{Q \le \tau_0\}]$ and $\mathbb{E}[XX^T 1\{Q > \tau_0\}]$ 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

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

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

Supplementary Materials

Online supplements are comprised of 6 appendices. In Appendix B, we provide the al- gorithm of constructing the confidence interval for τ_0. In Appendix C, we provide sufficient conditions for the identification 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. In Section F, we present the results of extensive Monte Carlo experiments, and Appendix G gives additional results for the empirical example.

Acknowledgment

The authors thank Bernd Fitzenberger, an editor, an associate editor, and three anonymous referees for helpful comments.

Funding

This work was supported in part by Promising-Pioneering Researcher Program through Seoul National University, by the European Research Council (ERC-2014-CoG-646917-ROMIA), by the Research and Scholarship Award grant of University of Maryland, and by the Australian Research Council (DP170100987).

References

Belloni, A., and Chernozhukov, V. (2011), "ℓ_1-Penalized Quantile Regression in High Dimensional Sparse Models," *Annals of Statistics*, 39, 82–130. [1184,1186,1187,1193]

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics*, 37, 1705–1732. [1185,1188]

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, 73, 325–349. [1184]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data, Methods, Theory and Applications*, New York: Springer. [1187,1192]

Callot, L., Caner, M., Kock, A. B., and Riquelme, J. A. (2017), "Sharp Threshold Detection Based on Sup-norm Error Rates in High-dimensional Models," *Journal of Business & Economic Statistics*, 35, 250–264. [1184]

Card, D., Mas, A., and Rothstein, J. (2008), "Tipping and the Dynamics of Segregation," *Quarterly Journal of Economics*, 123, 177–218. [1185,1191]

Chan, K.-S. (1993), "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *Annals of Statistics*, 21, 520–533. [1184,1193]

Chan, N. H., Ing, C.-K., Li, Y., and Yau, C. Y. (2017), "Threshold Estimation via Group Orthogonal Greedy Algorithm," *Journal of Business & Economic Statistics*, 35, 334–345. [1184]

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. [1184]

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, 77, 475–507. [1184,1192]

Ciuperca, G. (2013), "Quantile Regression in High-Dimension with Breaking," *Journal of Statistical Theory and Applications*, 12, 288–305. [1184]

Enikeeva, F., and Harchaoui, Z. (2013), "High-Dimensional Change-Point Detection with Sparse Alternatives," arXiv preprint, *http://arxiv.org/abs/1312.1900*. [1184]

Fan, J., Fan, Y., and Barut, E. (2014), "Adaptive Robust Variable Selection," *Annals of Statistics*, 42, 324–351. [1184]

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. [1186,1187]

Frick, K., Munk, A., and Sieling, H. (2014), "Multiscale Change Point Inference," *Journal of the Royal Statistical Society*, Series B, 76, 495–580. [1184]

Fryzlewicz, P. (2014), "Wild Binary Segmentation for Multiple Change-Point Detection," *Annals of Statistics*, 42, 2243–2281. [1192]

Hansen, B. E. (1996), "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430. [1184,1187]

——— (2000), "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–603. [1184,1193]

He, X., and Shao, Q.-M. (2000), "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 120–135. [1189]

Koenker, R. (2016), *quantreg: Quantile Regression, R Package Version 5.29*, CRAN, available at *https://cran.r-project.org/web/packages/quantreg/index.html*. [1187]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [1185]

Koenker, R., and Mizera, I. (2014), "Convex Optimization in R," *Journal of Statistical Software*, 60, 1–23. [1186]

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. [1184,1188]

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. [1188]

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. [1191]

——— (2016), "The Lasso for High Dimensional Regression with a Possible Change Point," *Journal of the Royal Statistical Society*, Series B, 78, 193–210. [1184,1185]

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*. [1192]

Li, D., and Ling, S. (2012), "On the Least Squares Estimation of Multiple-Regime Threshold Autoregressive Models," *Journal of Econometrics*, 167, 240–253. [1184,1188]

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*, eds. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Curran Associates, Inc., pp. 476–484. [1187]

Lovász, L., and Vempala, S. (2007), "The Geometry of Logconcave Functions and Sampling Algorithms," *Random Structures & Algorithms*, 30, 307–358. [1193]

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. [1184,1188]

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. [1188]

Seijo, E., and Sen, B. (2011a), "Change-Point in Stochastic Design Regression and the Bootstrap," *Annals of Statistics*, 39, 1580–1607. [1184]

——— (2011b), "A Continuous Mapping Theorem for the Smallest Argmax Functional," *Electronic Journal of Statistics*, 5, 421–439. [1184,1188]

Tong, H. (1990), *Non-Linear Time Series: A Dynamical System Approach*, Oxford: Oxford University Press. [1184]

van de Geer, S. A. (2008), "High-Dimensional Generalized Linear Models and the Lasso," *Annals of Statistics*, 36, 614–645. [1187]

Wang, L. (2013), "The L_1 Penalized LAD Estimator for High Dimensional Linear Regression," *Journal of Multivariate Analysis*, 120, 135–151. [1184]

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. [1184]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [1187]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimations in Non Concave Penalized Likelihood Models," *Annals of Statistics*, 36, 1509–1533. [1186]