The Annals of Statistics
2011, Vol. 39, No. 1, 82–130
DOI: 10.1214/10-AOS827

ℓ_1-PENALIZED QUANTILE REGRESSION IN HIGH-DIMENSIONAL SPARSE MODELS[1]

BY ALEXANDRE BELLONI AND VICTOR CHERNOZHUKOV

Duke University and Massachusetts Institute of Technology

We consider median regression and, more generally, a possibly infinite collection of quantile regressions in high-dimensional sparse models. In these models, the number of regressors p is very large, possibly larger than the sample size n, but only at most s regressors have a nonzero impact on each conditional quantile of the response variable, where s grows more slowly than n. Since ordinary quantile regression is not consistent in this case, we consider ℓ_1-penalized quantile regression (ℓ_1-QR), which penalizes the ℓ_1-norm of regression coefficients, as well as the post-penalized QR estimator (post-ℓ_1-QR), which applies ordinary QR to the model selected by ℓ_1-QR. First, we show that under general conditions ℓ_1-QR is consistent at the near-oracle rate $\sqrt{s/n}\sqrt{\log(p \vee n)}$, uniformly in the compact set $\mathcal{U} \subset (0, 1)$ of quantile indices. In deriving this result, we propose a partly pivotal, data-driven choice of the penalty level and show that it satisfies the requirements for achieving this rate. Second, we show that under similar conditions post-ℓ_1-QR is consistent at the near-oracle rate $\sqrt{s/n}\sqrt{\log(p \vee n)}$, uniformly over $\mathcal{U}$, even if the ℓ_1-QR-selected models miss some components of the true models, and the rate could be even closer to the oracle rate otherwise. Third, we characterize conditions under which ℓ_1-QR contains the true model as a submodel, and derive bounds on the dimension of the selected model, uniformly over $\mathcal{U}$; we also provide conditions under which hard-thresholding selects the minimal true model, uniformly over $\mathcal{U}$.

1. Introduction. Quantile regression is an important statistical method for analyzing the impact of regressors on the conditional distribution of a response variable (cf. [21, 23]). It captures the heterogeneous impact of regressors on different parts of the distribution [8], exhibits robustness to outliers [19], has excellent computational properties [28], and has wide applicability [19]. The asymptotic theory for quantile regression has been developed under both a fixed number of regressors and an increasing number of regressors. The asymptotic theory under a fixed number of regressors is given in [13, 15, 17, 21, 27] and others. The asymptotic theory under an increasing number of regressors is given in [16] and [1, 4], covering the case where the number of regressors p is negligible relative to the sample size n [i.e., $p = o(n)$].

Received December 2009; revised April 2010.
[1]Supported by the NSF Grant SES-0752266.
AMS 2000 subject classifications. Primary 62H12, 62J99; secondary 62J07.
Key words and phrases. Median regression, quantile regression, sparse models.

In this paper, we consider quantile regression in high-dimensional sparse models (HDSMs). In such models, the overall number of regressors p is very large, possibly much larger than the sample size n. However, the number of significant regressors for each conditional quantile of interest is at most s, which is smaller than the sample size, that is, $s = o(n)$. HDSMs [7, 12, 26] have emerged to deal with many new applications arising in biometrics, signal processing, machine learning, econometrics, and other areas of data analysis where high-dimensional data sets have become widely available.

A number of papers have begun to investigate estimation of HDSMs, focusing primarily on penalized mean regression, with the ℓ_1-norm acting as a penalty function [7, 12, 22, 26, 32, 34]. References [7, 12, 22, 26, 34] demonstrated the fundamental result that ℓ_1-penalized least squares estimators achieve the rate $\sqrt{s/n}\sqrt{\log p}$, which is very close to the oracle rate $\sqrt{s/n}$ achievable when the true model is known. Reference [32] demonstrated a similar fundamental result on the excess forecasting error loss under both quadratic and nonquadratic loss functions. Thus, the estimator can be consistent and can have excellent forecasting performance even under very rapid, nearly exponential, growth of the total number of regressors p. See [7, 9–11, 14, 25, 29] for many other interesting developments and a detailed review of the existing literature.

Our paper's contribution is to develop a set of results on model selection and rates of convergence for quantile regression within the HDSM framework. Since ordinary quantile regression is inconsistent in HDSMs, we consider quantile regression penalized by the ℓ_1-norm of parameter coefficients, denoted ℓ_1-QR. First, we show that under general conditions ℓ_1-QR estimates of regression coefficients and regression functions are consistent at the near-oracle rate $\sqrt{s/n}\sqrt{\log(p \vee n)}$, uniformly in a compact interval $\mathcal{U} \subset (0, 1)$ of quantile indices.[2] (This result is different from, and hence complementary to [32]'s fundamental results on the rates for excess forecasting error loss.) Second, in order to make ℓ_1-QR practical, we propose a partly pivotal, data-driven choice of the penalty level, and show that this choice leads to the same sharp convergence rate. Third, we show that ℓ_1-QR correctly selects the true model as a valid submodel when the nonzero coefficients of the true model are well separated from zero. Fourth, we also propose and analyze the post-penalized estimator (post-ℓ_1-QR), which applies ordinary, unpenalized quantile regression to the model selected by the penalized estimator, and thus aims at reducing the regularization bias of the penalized estimator. We show that under similar conditions post-ℓ_1-QR can perform as well as ℓ_1-QR in terms of the rate of convergence, uniformly over $\mathcal{U}$, even if the ℓ_1-QR-based model selection misses some components of the true models. This occurs because ℓ_1-QR-based model selection only misses those components that have relatively small coefficients. Moreover, post-ℓ_1-QR can perform better than ℓ_1-QR if the ℓ_1-QR-based

[2]Under $s \to \infty$, the oracle rate, uniformly over a proper compact interval $\mathcal{U}$, is $\sqrt{(s/n)\log n}$, cf. [4]; the oracle rate for a single quantile index is $\sqrt{s/n}$; cf. [16].

model selection correctly includes all components of the true model as a subset. (Obviously, post-ℓ_1-QR can perform as well as the oracle if the ℓ_1-QR perfectly selects the true model, which is, however, unrealistic for many designs of interest.) Fifth, we illustrate the use of ℓ_1-QR and post-ℓ_1-QR with a Monte Carlo experiment. To the best of our knowledge, all of the above results are new and contribute to the literature on HDSMs. Our results on post-penalized estimators and some proof techniques could also be of interest in other problems. We provide further technical comparisons to the literature in Section 2.

1.1. *Notation.* In what follows, we implicitly index all parameter values by the sample size n, but we omit the index whenever this does not cause confusion. We use the empirical process notation as defined in [33]. In particular, given a random sample $Z_1, \ldots, Z_n$, let $\mathbb{G}_n(f) = \mathbb{G}_n(f(Z_i)) := n^{-1/2} \sum_{i=1}^n (f(Z_i) - \mathrm{E}[f(Z_i)])$ and $\mathbb{E}_n f = \mathbb{E}_n f(Z_i) := \sum_{i=1}^n f(Z_i)/n$. We use the notation $a \lesssim b$ to denote $a = O(b)$, that is, $a \leq cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. We also use the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We denote the ℓ_2-norm by $\|\cdot\|$, ℓ_1-norm by $\|\cdot\|_1$, ℓ_∞-norm by $\|\cdot\|_\infty$ and the ℓ_0-"norm" by $\|\cdot\|_0$ (i.e., the number of nonzero components). We denote by $\|\beta\|_{1,n} = \sum_{j=1}^p \widehat{\sigma}_j |\beta_j|$ the ℓ_1-norm weighted by $\widehat{\sigma}_j$'s. Finally, given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$.

2. The estimator, the penalty level and overview of rate results. In this section, we formulate the setting and the estimator, and state primitive regularity conditions. We also provide an overview of the main results.

2.1. *Basic setting.* The setting of interest corresponds to a parametric quantile regression model, where the dimension p of the underlying model increases with the sample size n. Namely, we consider a response variable y and p-dimensional covariates x such that the uth conditional quantile function of y given x is given by

$$F^{-1}_{y_i|x_i}(u|x_i) = x'\beta(u), \qquad \beta(u) \in \mathbb{R}^p \qquad \text{for all } u \in \mathcal{U}, \tag{2.1}$$

where $\mathcal{U} \subset (0, 1)$ is a compact set of quantile indices. Recall that the uth conditional quantile $F^{-1}_{y_i|x_i}(u|x)$ is the inverse of the conditional distribution function $F_{y_i|x_i}(y|x_i)$ of y_i given x_i. We consider the case where the dimension p of the model is large, possibly much larger than the available sample size n, but the true model $\beta(u)$ has a sparse support

$$T_u = \mathrm{support}(\beta(u)) = \big\{j \in \{1, \ldots, p\} : |\beta_j(u)| > 0\big\}$$

having only $s_u \leq s \leq n/\log(n \vee p)$ nonzero components for all $u \in \mathcal{U}$.

The population coefficient $\beta(u)$ is known to minimize the criterion function

$$Q_u(\beta) = \mathrm{E}[\rho_u(y - x'\beta)], \tag{2.2}$$

where $\rho_u(t) = (u - 1\{t \le 0\})t$ is the asymmetric absolute deviation function [21]. Given a random sample $(y_1, x_1), \ldots, (y_n, x_n)$, the quantile regression estimator of $\beta(u)$ is defined as a minimizer of the empirical analog of (2.2):

$$\widehat{Q}_u(\beta) = \mathbb{E}_n[\rho_u(y_i - x_i'\beta)]. \tag{2.3}$$

In high-dimensional settings, particularly when $p \ge n$, ordinary quantile regression is generally inconsistent, which motivates the use of penalization in order to remove all, or at least nearly all, regressors whose population coefficients are zero, thereby possibly restoring consistency. A penalization that has proven quite useful in least squares settings is the ℓ_1-penalty leading to the Lasso estimator [30].

2.2. *Penalized and post-penalized estimators.* The ℓ_1-penalized quantile regression estimator $\widehat{\beta}(u)$ is a solution to the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \widehat{Q}_u(\beta) + \frac{\lambda\sqrt{u(1-u)}}{n} \sum_{j=1}^{p} \widehat{\sigma}_j |\beta_j|, \tag{2.4}$$

where $\widehat{\sigma}_j^2 = \mathbb{E}_n[x_{ij}^2]$. The criterion function in (2.4) is the sum of the criterion function (2.3) and a penalty function given by a scaled ℓ_1-norm of the parameter vector. The overall penalty level $\lambda\sqrt{u(1-u)}$ depends on each quantile index u, while λ will depend on the set $\mathcal{U}$ of quantile indices of interest. The ℓ_1-penalized quantile regression has been considered in [18] under small (fixed) p asymptotics. It is important to note that the penalized quantile regression problem (2.4) is equivalent to a linear programming problem (see Appendix C) with a dual version that is useful for analyzing the sparsity of the solution. When the solution is not unique, we define $\widehat{\beta}(u)$ as any optimal basic feasible solution (see, e.g., [6]). Therefore, the problem (2.4) can be solved in polynomial time, avoiding the computational curse of dimensionality. Our goal is to derive the rate of convergence and model selection properties of this estimator.

The post-penalized estimator (post-ℓ_1-QR) applies ordinary quantile regression to the model $\widehat{T}_u$ selected by the ℓ_1-penalized quantile regression. Specifically, set

$$\widehat{T}_u = \mathrm{support}(\widehat{\beta}(u)) = \{j \in \{1, \ldots, p\} : |\widehat{\beta}_j(u)| > 0\},$$

and define the post-penalized estimator $\widetilde{\beta}(u)$ as

$$\widetilde{\beta}(u) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p : \beta_{\widehat{T}_u^c} = 0} \widehat{Q}_u(\beta), \tag{2.5}$$

which removes from further estimation the regressors that were not selected. If the model selection works perfectly—that is, $\widehat{T}_u = T_u$—then this estimator is simply the oracle estimator, whose properties are well known. However, perfect model

selection might be unlikely for many designs of interest. Rather, we are interested in the more realistic scenario where the first-step estimator $\widehat{\beta}(u)$ fails to select some components of $\beta(u)$. Our goal is to derive the rate of convergence for the post-penalized estimator and show it can perform well under this scenario.

2.3. *The choice of the penalty level* λ. In order to describe our choice of the penalty level λ, we introduce the random variable

$$\Lambda = n \sup_{u \in \mathcal{U}} \max_{1 \le j \le p} \left| \mathbb{E}_n \left[\frac{x_{ij}(u - 1\{u_i \le u\})}{\widehat{\sigma}_j \sqrt{u(1-u)}} \right] \right|, \tag{2.6}$$

where $u_1, \ldots, u_n$ are i.i.d. uniform $(0, 1)$ random variables, independently distributed from the regressors, $x_1, \ldots, x_n$. The random variable Λ has a known, that is, pivotal, distribution conditional on $X = [x_1, \ldots, x_n]'$. We then set

$$\lambda = c \cdot \Lambda(1 - \alpha | X), \tag{2.7}$$

where $\Lambda(1 - \alpha|X) :=$ $(1 - \alpha)$-quantile of Λ conditional on X, and the constant $c > 1$ depends on the design.[3] Thus, the penalty level depends on the pivotal quantity $\Lambda(1 - \alpha|X)$ and the design. Under assumptions D.1–D.4, we can set $c = 2$, similar to [7]'s choice for least squares. Furthermore, we recommend computing $\Lambda(1 - \alpha|X)$ using simulation of Λ.[4] Our concrete recommendation for practice is to set $1 - \alpha = 0.9$.

The parameter $1 - \alpha$ is the confidence level in the sense that, as in [7], our (nonasymptotic) bounds on the estimation error will contract at the optimal rate with this probability. We refer the reader to Koenker [20] for an implementation of our choice of penalty level and practical suggestions concerning the confidence level. In particular, both here and in Koenker [20], the confidence level $1 - \alpha = 0.9$ gave good performance results in terms of balancing regularization bias with estimation variance. Cross-validation may also be used to choose the confidence level $1 - \alpha$. Finally, we should note that, as in [7], our theoretical bounds allow for any choice of $1 - \alpha$ and are stated as a function of $1 - \alpha$.

The formal rationale behind the choice (2.7) for the penalty level λ is that this choice leads precisely to the optimal rates of convergence for ℓ_1-QR. (The same or slightly higher choice of λ also guarantees good performance of post-ℓ_1-QR.) Our general strategy for choosing λ follows [7], who recommend selecting λ so that it dominates a relevant measure of noise in the sample criterion function, specifically the supremum norm of a suitably rescaled gradient of the sample criterion function evaluated at the true parameter value. In our case,

[3] c depends only on the constant c_0 appearing in condition D.4; when $c_0 \ge 9$, it suffices to set $c = 2$.

[4] We also provide analytical bounds on $\Lambda(1 - \alpha|X)$ of the form $C(\alpha, \mathcal{U})\sqrt{n \log p}$ for some numeric constant $C(\alpha, \mathcal{U})$. We recommend simulation because it accounts for correlation among the columns of X in the sample.

this general strategy leads precisely to the choice (2.7). Indeed, a (sub)gradient $\widehat{S}_u(\beta(u)) = \mathbb{E}_n[(u - 1\{y_i \le x_i'\beta(u)\})x_i] \in \partial \widehat{Q}_u(\beta(u))$ of the quantile regression objective function evaluated at the truth has a pivotal representation, namely $\widehat{S}_u(\beta(u)) = \mathbb{E}_n[(u - 1\{u_i \le u\})x_i]$ for $u_1, \ldots, u_n$ i.i.d. uniform $(0, 1)$ conditional on X, and so we can represent Λ as in (2.6), and, thus, choose λ as in (2.7).

2.4. *General regularity conditions*. We consider the following conditions on a sequence of models indexed by n with parameter dimension $p = p_n \to \infty$. In these conditions, all constants can depend on n, but we omit the explicit indexing by n to ease exposition.

D.1 (*Sampling and smoothness*). Data $(y_i, x_i')'$, $i = 1, \ldots, n$, are an i.i.d. sequence of real $(1 + p)$-vectors, with the conditional u-quantile function given by (2.1) for each $u \in \mathcal{U}$, with the first component of x_i equal to one, and $n \wedge p \ge 3$. For each value x in the support of x_i, the conditional density $f_{y_i|x_i}(y|x)$ is continuously differentiable in y at each $y \in \mathbb{R}$, and $f_{y_i|x_i}(y|x)$ and $\frac{\partial}{\partial y} f_{y_i|x_i}(y|x)$ are bounded in absolute value by constants $\bar{f}$ and $\bar{f}'$, uniformly in $y \in \mathbb{R}$ and x in the support x_i. Moreover, the conditional density of y_i evaluated at the conditional quantile $x_i'\beta(u)$ is bounded away from zero uniformly in $\mathcal{U}$, that is, $f_{y_i|x_i}(x'\beta(u)|x) > \underline{f} > 0$ uniformly in $u \in \mathcal{U}$ and x in the support of x_i.

Condition D.1 imposes only mild smoothness assumptions on the conditional density of the response variable given regressors, and does not impose any normality or homoscedasticity assumptions. The assumption that the conditional density is bounded below at the conditional quantile is standard, but we can replace it by the slightly more general condition $\inf_{u\in\mathcal{U}} \inf_{\delta\neq 0}(\delta' J_u \delta)/(\delta' \mathrm{E}[x_i x_i']\delta) \ge \underline{f} > 0$, on the Jacobian matrices

$$J_u = \mathrm{E}[f_{y_i|x_i}(x_i'\beta(u)|x_i)x_i x_i'] \qquad \text{for all } u \in \mathcal{U},$$

throughout the paper; see [3] for a further generalization.

D.2 [*Sparsity and smoothness of* $u \mapsto \beta(u)$]. Let $\mathcal{U}$ be a compact subset of $(0, 1)$. The coefficients $\beta(u)$ in (2.1) are sparse and smooth with respect to $u \in \mathcal{U}$:

$$\sup_{u\in\mathcal{U}} \|\beta(u)\|_0 \le s \quad \text{and} \quad \|\beta(u) - \beta(u')\| \le L|u - u'| \qquad \text{for all } u, u' \in \mathcal{U},$$

where $s \ge 1$, and $\log L \le C_L \log(p \vee n)$ for some constant C_L.

Condition D.2 imposes sparsity and smoothness on the behavior of the quantile regression coefficients $\beta(u)$ as we vary the quantile index u.

D.3 (*Well-behaved covariates*). Covariates are normalized such that $\sigma_j^2 = \mathrm{E}[x_{ij}^2] = 1$ for all $j = 1, \ldots, p$, and $\widehat{\sigma}_j^2 = \mathbb{E}_n[x_{ij}^2]$ obeys $P(\max_{1\le j\le p}|\widehat{\sigma}_j - 1| \le 1/2) \ge 1 - \gamma \to 1$ as $n \to \infty$.

Condition D.3 requires that $\widehat{\sigma}_j$ does not deviate too much from σ_j and normalizes $\sigma_j^2 = 1$.

In order to state the next assumption, for some $c_0 \geq 0$ and each $u \in \mathcal{U}$, define

$$A_u := \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leq c_0 \|\delta_{T_u}\|_1, \|\delta_{T_u^c}\|_0 \leq n\},$$

which will be referred to as the restricted set. Define $\overline{T}_u(\delta, m) \subset \{1, \ldots, p\} \setminus T_u$ as the support of the m largest in absolute value components of the vector δ outside of $T_u = \text{support}(\beta(u))$, where $\overline{T}_u(\delta, m)$ is the empty set if $m = 0$.

D.4 (*Restricted identifiability and nonlinearity*). For some constants $m \geq 0$ and $c_0 \geq 9$, the matrix $\mathrm{E}[x_i x_i']$ satisfies

$$\text{(RE}(c_0, m)\text{)} \qquad \kappa_m^2 := \inf_{u \in \mathcal{U}} \inf_{\delta \in A_u, \delta \neq 0} \frac{\delta' \mathrm{E}[x_i x_i'] \delta}{\|\delta_{T_u \cup \overline{T}_u(\delta, m)}\|^2} > 0$$

and $\log(\underline{f} \kappa_0^2) \leq C_f \log(n \vee p)$ for some constant C_f. Moreover,

$$\text{(RNI}(c_0)\text{)} \qquad q := \frac{3}{8} \frac{\underline{f}^{3/2}}{\bar{f}'} \inf_{u \in \mathcal{U}} \inf_{\delta \in A_u, \delta \neq 0} \frac{\mathrm{E}[|x_i'\delta|^2]^{3/2}}{\mathrm{E}[|x_i'\delta|^3]} > 0.$$

The restricted eigenvalue (RE) condition is analogous to the condition in [7] and [12]; see [7] and [12] for different sufficient primitive conditions that yield bounds on κ_m. Also, since κ_m is nonincreasing in m, (RE(c_0, m)) for any $m > 0$ implies (RE$(c_0, 0)$). The restricted nonlinear impact (RNI) coefficient q appearing in D.4 is a new concept, which controls the quality of minoration of the quantile regression objective function by a quadratic function over the restricted set.

Finally, we state another condition needed to derive results on the post-model selected estimator. In order to state the condition, define the sparse set $\widetilde{A}_u(\widetilde{m}) = \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_0 \leq \widetilde{m}\}$ for $\widetilde{m} \geq 0$ and $u \in \mathcal{U}$.

D.5 (*Sparse identifiability and nonlinearity*). The matrix $\mathrm{E}[x_i x_i']$ satisfies for some $\widetilde{m} \geq 0$

$$\text{(SE}(\widetilde{m})\text{)} \qquad \widetilde{\kappa}_{\widetilde{m}}^2 := \inf_{u \in \mathcal{U}} \inf_{\delta \in \widetilde{A}_u(\widetilde{m}), \delta \neq 0} \frac{\delta' \mathrm{E}[x_i x_i'] \delta}{\delta' \delta} > 0$$

and

$$\text{(SNI}(\widetilde{m})\text{)} \qquad \widetilde{q}_{\widetilde{m}} := \frac{3}{8} \frac{\underline{f}^{3/2}}{\bar{f}'} \inf_{u \in \mathcal{U}} \inf_{\delta \in \widetilde{A}_u(\widetilde{m}), \delta \neq 0} \frac{\mathrm{E}[|x_i'\delta|^2]^{3/2}}{\mathrm{E}[|x_i'\delta|^3]} > 0.$$

We invoke the sparse eigenvalue (SE) condition in order to analyze the post-penalized estimator (2.5). This assumption is similar to the conditions used in [26] and [34] to analyze Lasso. Our form of the (SE) condition is neither less nor more general than the (RE) condition. The SNI coefficient $\widetilde{q}_{\widetilde{m}}$ controls the quality of minoration of the quantile regression objective function by a quadratic function over sparse neighborhoods of the true parameter.

2.5. *Examples of simple sufficient conditions.* In order to highlight the nature and usefulness of conditions D.1–D.5, it is instructive to state some simple sufficient conditions (note that D.1–D.5 allow for much more general conditions). We relegate the proofs of this section to [2] for brevity.

DESIGN 1 (Location model with correlated normal design). Let us consider estimating a standard location model

$$y = x'\beta^o + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, $\sigma > 0$ is fixed, $x = (1, z')'$, with $z \sim N(0, \Sigma)$, where Σ has ones in the diagonal, a minimum eigenvalue bounded away from zero by a constant $\kappa^2 > 0$, and a maximum eigenvalue bounded from above, uniformly in n.

LEMMA 1. *Under Design 1 with* $\mathcal{U} = [\xi, 1-\xi]$, $\xi > 0$, *conditions* D.1–D.5 *are satisfied with*

$$\bar{f} = 1/[\sqrt{2\pi}\sigma], \qquad \bar{f}' = \sqrt{e/[2\pi]}/\sigma^2, \qquad \underline{f} = 1/\sqrt{2\pi\xi}\sigma,$$
$$\|\beta(u)\|_0 \le \|\beta^o\|_0 + 1, \qquad \gamma = 2p\exp(-n/24), \qquad L = \sigma/\xi,$$
$$\kappa_m \wedge \widetilde{\kappa}_{\widetilde{m}} \ge \kappa, \qquad q \wedge \widetilde{q}_{\widetilde{m}} \ge (3/[32\xi^{3/4}])\sqrt{\sqrt{2\pi}\sigma/e}.$$

Note that the normality of errors can be easily relaxed by allowing for the disturbance ε to have a smooth density that obeys the conditions stated in D.1. The conditions on the population design matrix can also be replaced by more general primitive conditions specified in Remark 2.1.

DESIGN 2 (Location-scale model with bounded regressors). Let us consider estimating a standard location-scale model

$$y = x'\beta^o + x'\eta \cdot \varepsilon,$$

where $\varepsilon \sim F$ independent of x, with a continuously differentiable probability density function f. We assume that the population design matrix $\mathrm{E}[xx']$ has ones in the diagonal and has eigenvalues uniformly bounded away from zero and from above, $x_1 = 1$, $\max_{1\le j\le p}|x_j| \le K_B$. Moreover, the vector η is such that $0 < \upsilon \le x'\eta \le \Upsilon < \infty$ for all values of x.

LEMMA 2. *Under Design 2 with* $\mathcal{U} = [\xi, 1-\xi]$, $\xi > 0$, *conditions* D.1–D.5 *are satisfied with*

$$\bar{f} \le \max_{\varepsilon} f(\varepsilon)/\upsilon, \qquad \bar{f}' \le \max_{\varepsilon} f'(\varepsilon)/\upsilon^2,$$
$$\underline{f} = \min_{u\in\mathcal{U}} f(F^{-1}(u))/\Upsilon, \qquad \|\beta(u)\|_0 \le \|\beta^o\|_0 + \|\eta\|_0 + 1,$$

$$\gamma = 2p\exp(-n/[8K_B^4]), \qquad \kappa_m \wedge \widetilde{\kappa}_{\widetilde{m}} \geq \kappa, \qquad L = \|\eta\|\underline{f},$$
$$q \geq \frac{3}{8}\frac{\underline{f}^{3/2}}{\bar{f}'}\kappa/[10K_B\sqrt{s}], \qquad \widetilde{q}_{\widetilde{m}} \geq \frac{3}{8}\frac{\underline{f}^{3/2}}{\bar{f}'}\kappa/[K_B\sqrt{s+\widetilde{m}}].$$

COMMENT 2.1 (Conditions on $\mathrm{E}[x_i x_i']$). The conditions on the population design matrix can also be replaced by more general primitive conditions of the form stated in [7] and [12]. For example, conditions on sparse eigenvalues suffice as shown in [7]. Denote the minimum and maximum eigenvalue of the population design matrix by

$$\varphi_{\min}(m) = \min_{\|\delta\|=1, \|\delta\|_0 \leq m} \frac{\delta'\mathrm{E}[x_i x_i']\delta}{\delta'\delta} \quad \text{and}$$
$$\varphi_{\max}(m) = \max_{\|\delta\|=1, \|\delta\|_0 \leq m} \frac{\delta'\mathrm{E}[x_i x_i']\delta}{\delta'\delta}. \tag{2.8}$$

Assuming that for some $m \geq s$ we have $m\varphi_{\min}(m+s) \geq c_0^2 s\varphi_{\max}(m)$, then

$$\kappa_m \geq \sqrt{\varphi_{\min}(s+m)}\big(1 - c_0\sqrt{s\varphi_{\max}(s)/[m\varphi_{\min}(s+m)]}\big)$$

and

$$\widetilde{\kappa}_{\widetilde{m}} \geq \varphi_{\min}(s+m).$$

2.6. *Overview of main results.* Here, we discuss our results under the simple setup of Design 1 and under $1/p \leq \alpha \to 0$ and $\gamma \to 0$. These simple assumptions allow us to straightforwardly compare our rate results to those obtained in the literature. We state our more general nonasymptotic results under general conditions in the subsequent sections. Our first main rate result is that ℓ_1-QR, with our choice (2.7) of parameter λ, satisfies

$$\sup_{u \in \mathcal{U}} \|\widehat{\beta}(u) - \beta(u)\| \lesssim_P \frac{1}{\underline{f}\kappa_0\kappa_s}\sqrt{\frac{s\log(n \vee p)}{n}}, \tag{2.9}$$

provided that the upper bound on the number of nonzero components s satisfies

$$\frac{\sqrt{s\log(n \vee p)}}{\sqrt{n}\underline{f}^{1/2}\kappa_0 q} \to 0. \tag{2.10}$$

Note that κ_0, κ_s, $\underline{f}$ and q are bounded away from zero in this example. Therefore, the rate of convergence is $\sqrt{s/n} \cdot \sqrt{\log(n \vee p)}$ uniformly in the set of quantile indices $u \in \mathcal{U}$, which is very close to the oracle rate when p grows polynomially in n. Further, we note that our resulting restriction (2.10) on the dimension s of the true models is very weak; when p is polynomial in n, s can be of almost the same order as n, namely $s = o(n/\log n)$.

Our second main result is that the dimension $\|\widehat{\beta}(u)\|_0$ of the model selected by the ℓ_1-penalized estimator is of the same stochastic order as the dimension s of the true models, namely

$$\sup_{u\in\mathcal{U}} \|\widehat{\beta}(u)\|_0 \lesssim_P s. \tag{2.11}$$

Further, if the parameter values of the minimal true model are well separated from zero, then with a high probability the model selected by the ℓ_1-penalized estimator correctly nests the true minimal model:

$$T_u = \operatorname{support}(\beta(u)) \subseteq \widehat{T}_u = \operatorname{support}(\widehat{\beta}(u)) \qquad \text{for all } u \in \mathcal{U}. \tag{2.12}$$

Moreover, we provide conditions under which a hard-thresholded version of the estimator selects the correct support.

Our third main result is that the post-penalized estimator, which applies ordinary quantile regression to the selected model, obeys

$$\begin{aligned}\sup_{u\in\mathcal{U}} \|\widetilde{\beta}(u) - \beta(u)\| \lesssim_P {} & \frac{1}{\underline{f}\widetilde{\kappa}_{\widehat{m}}^2}\sqrt{\frac{\widehat{m}\log(n\vee p) + s\log n}{n}} \\ & + \frac{\sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\}}{\underline{f}\kappa_0\widetilde{\kappa}_{\widehat{m}}}\sqrt{\frac{s\log(n\vee p)}{n}},\end{aligned} \tag{2.13}$$

where $\widehat{m} = \sup_{u\in\mathcal{U}} \|\widehat{\beta}_{T_u^c}(u)\|_0$ is the maximum number of wrong components selected for any quantile index $u \in \mathcal{U}$, provided that the bound on the number of nonzero components s obeys the growth condition (2.10) and

$$\frac{\sqrt{\widehat{m}\log(n\vee p) + s\log n}}{\sqrt{n}\,\underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}\widetilde{q}_{\widehat{m}}} \to_P 0. \tag{2.14}$$

[Note that when $\mathcal{U}$ is a singleton, the $s\log n$ factor in (2.13) becomes s.]

We see from (2.13) that post-ℓ_1-QR can perform well in terms of the rate of convergence even if the selected model $\widehat{T}_u$ fails to contain the true model T_u. Indeed, since in this design $\widehat{m} \lesssim_P s$, post-ℓ_1-QR has the rate of convergence $\sqrt{s/n}\cdot\sqrt{\log(n\vee p)}$, which is the same as the rate of convergence of ℓ_1-QR. The intuition for this result is that the ℓ_1-QR based model selection can only miss covariates with relatively small coefficients, which then permits post-ℓ_1-QR to perform as well or even better due to reductions in bias, as confirmed by our computational experiments.

We also see from (2.13) that post-ℓ_1-QR can perform better than ℓ_1-QR in terms of the rate of convergence if the number of wrong components selected obeys $\widehat{m} = o_P(s)$ and the selected model contains the true model, $\{T_u \subseteq \widehat{T}_u\}$ with probability converging to one. In this case, post-ℓ_1-QR has the rate of convergence $\sqrt{(o_P(s)/n)\log(n\vee p) + (s/n)\log n}$, which is faster than the rate of convergence of ℓ_1-QR. In the extreme case of perfect model selection, that is, when $\widehat{m} = 0$, the

rate of post-ℓ_1-QR becomes $\sqrt{(s/n)\log n}$ uniformly in $\mathcal{U}$. (When $\mathcal{U}$ is a singleton, the $\log n$ factor drops out.) Note that inclusion $\{T_u \subseteq \widehat{T}_u\}$ necessarily happens when the coefficients of the true models are well separated from zero, as we stated above. Note also that the condition $\widehat{m} = o(s)$ or even $\widehat{m} = 0$ could occur under additional conditions on the regressors (such as the mutual coherence conditions that restrict the maximal pairwise correlation of regressors). Finally, we note that our second restriction (2.14) on the dimension s of the true models is very weak in this design; when p is polynomial in n, s can be of almost the same order as n, namely $s = o(n/\log n)$.

To the best of our knowledge, all of the results presented above are new, both for the single ℓ_1-penalized quantile regression problem as well as for the infinite collection of ℓ_1-penalized quantile regression problems. These results therefore contribute to the rate results obtained for ℓ_1-penalized mean regression and related estimators in the fundamental papers of [7, 12, 22, 26, 32, 34]. The results on post-ℓ_1 penalized quantile regression had no analogs in the literature on mean regression, apart from the rather exceptional case of perfect model selection, in which case the post-penalized estimator is simply the oracle. Building on the current work these results have been extended to mean regression in [5]. Our results on the sparsity of ℓ_1-QR and model selection also contribute to the analogous results for mean regression [26]. Also, our rate results for ℓ_1-QR are different from, and hence complementary to, the fundamental results in [32] on the excess forecasting loss under possibly nonquadratic loss functions, which also specializes the results to density estimation, mean regression, and logistic regression. In principle, we could apply theorems in [32] to the single quantile regression problem to derive the bounds on the excess loss $\mathrm{E}[\rho_u(y_i - x_i'\widehat{\beta}(u))] - \mathrm{E}[\rho_u(y_i - x_i'\beta(u))]$.[5] However, these bounds would not imply our results (2.7), (2.9), (2.11), (2.12) and (2.13), which characterize the rates of estimating coefficients $\beta(u)$ by ℓ_1-QR and post-ℓ_1-QR, sparsity and model selection properties, and the data-driven choice of the penalty level.

3. Main results and main proofs. In this section, we derive rates of convergence for ℓ_1-QR and post-ℓ_1-QR, sparsity bounds, and model selection results.

3.1. *Bounds on* $\Lambda(1-\alpha|X)$. We start with a characterization of Λ and its $(1-\alpha)$-quantile, $\Lambda(1-\alpha|X)$, which determines the magnitude of our suggested penalty level λ via equation (2.7).

[5]Of course, such a derivation would entail some difficult work, since we must verify some high-level assumptions made directly on the performance of the oracle and penalized estimators in population and others (cf. [32], conditions I.1 and I.2, where I.2 assumes uniform in x_i consistency of the penalized estimator in the population, and does not hold in our main examples, e.g., in Design 1 with normal regressors).

THEOREM 1 [Bounds on $\Lambda(1-\alpha|X)$]. *Let* $W_{\mathcal{U}} = \max_{u\in\mathcal{U}} 1/\sqrt{u(1-u)}$. *There is a universal constant* C_Λ *such that*:

(i) $P(\Lambda \geq k \cdot C_\Lambda W_{\mathcal{U}} \sqrt{n \log p}|X) \leq p^{-k^2+1}$,
(ii) $\Lambda(1-\alpha|X) \leq \sqrt{1+\log(1/\alpha)/\log p} \cdot C_\Lambda W_{\mathcal{U}} \sqrt{n \log p}$ *with probability* 1.

3.2. *Rates of convergence.* In this section, we establish the rate of convergence of ℓ_1-QR. We start with the following preliminary result which shows that if the penalty level exceeds the specified threshold, for each $u \in \mathcal{U}$, the estimator $\widehat{\beta}(u) - \beta(u)$ will belong to the restricted set $A_u := \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leq c_0 \|\delta_{T_u}\|_1, \|\delta_{T_u^c}\|_0 \leq n\}$.

LEMMA 3 (Restricted set). 1. *Under* D.3, *with probability at least* $1-\gamma$ *we have for every* $\delta \in \mathbb{R}^p$ *that*

$$\tfrac{2}{3}\|\delta\|_{1,n} \leq \|\delta\|_1 \leq 2\|\delta\|_{1,n}. \tag{3.1}$$

2. *Moreover, if for some* $\alpha \in (0,1)$

$$\lambda \geq \lambda_0 := \frac{c_0+3}{c_0-3}\Lambda(1-\alpha|X), \tag{3.2}$$

then with probability at least $1-\alpha-\gamma$, *uniformly in* $u \in \mathcal{U}$, *we have* (3.1) *and*

$$\widehat{\beta}(u) - \beta(u) \in A_u = \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_1 \leq c_0 \|\delta_{T_u}\|_1, \|\delta_{T_u^c}\|_0 \leq n\}.$$

This result is inspired by the analogous result for least squares in [7].

LEMMA 4 (Identifiability relations over restricted set). *Condition* D.4, *namely* (RE(c_0, m)) *and* (RNI(c_0)), *implies that for any* $\delta \in A_u$ *and* $u \in \mathcal{U}$,

$$\|(\mathrm{E}[x_i x_i'])^{1/2}\delta\| \leq \|J_u^{1/2}\delta\|/\underline{f}^{1/2}, \tag{3.3}$$

$$\|\delta_{T_u}\|_1 \leq \sqrt{s}\|J_u^{1/2}\delta\|/[\underline{f}^{1/2}\kappa_0], \tag{3.4}$$

$$\|\delta\|_1 \leq \sqrt{s}(1+c_0)\|J_u^{1/2}\delta\|/[\underline{f}^{1/2}\kappa_0], \tag{3.5}$$

$$\|\delta\| \leq \big(1+c_0\sqrt{s/m}\big)\|J_u^{1/2}\delta\|/[\underline{f}^{1/2}\kappa_m], \tag{3.6}$$

$$Q_u\big(\beta(u)+\delta\big) - Q_u(\beta(u)) \geq (\|J_u^{1/2}\delta\|^2/4) \wedge (q\|J_u^{1/2}\delta\|). \tag{3.7}$$

This second preliminary result derives identifiability relations over A_u. It shows that the coefficients $\underline{f}$, κ_0 and κ_m control moduli of continuity between various norms over the restricted set A_u, and the RNI coefficient q controls the quality of minoration of the objective function by a quadratic function over A_u.

Finally, the third preliminary result derives bounds on the empirical error over A_u.

LEMMA 5 (Control of empirical error). *Under* D.1–D.4, *for any* $t > 0$ *let*

$$\epsilon(t) := \sup_{u\in\mathcal{U},\delta\in A_u,\|J_u^{1/2}\delta\|\le t} \big|\widehat{Q}_u(\beta(u)+\delta) - Q_u(\beta(u)+\delta) - \big(\widehat{Q}_u(\beta(u)) - Q_u(\beta(u))\big)\big|.$$

Then, there is a universal constant C_E *such that for any* $A > 1$, *with probability at least* $1 - 3\gamma - 3p^{-A^2}$

$$\epsilon(t) \le t \cdot C_E \cdot \frac{(1+c_0)A}{\underline{f}^{1/2}\kappa_0}\sqrt{\frac{s\log(p \vee [L\underline{f}^{1/2}\kappa_0/t])}{n}}.$$

In order to prove the lemma, we use a combination of chaining arguments and exponential inequalities for contractions [24]. Our use of the contraction principle is inspired by its fundamentally innovative use in [32]; however, the use of the contraction principle alone is not sufficient in our case. Indeed, first we need to make some adjustments to obtain error bounds over the neighborhoods defined by the intrinsic norm $\|J_u^{1/2}\cdot\|$ instead of the $\|\cdot\|_1$ norm; and second, we need to use chaining over $u \in \mathcal{U}$ to get uniformity over $\mathcal{U}$.

Armed with Lemmas 3–5, we establish the first main result. The result depends on the constants C_Λ, C_E, C_L and C_f defined in Theorem 1, Lemma 5, conditions D.2 and D.4.

THEOREM 2 (Uniform bounds on estimation error of ℓ_1-QR). *Assume conditions* D.1–D.4 *hold, and let*

$$C > 2C_\Lambda\sqrt{1+\log(1/\alpha)/\log p} \vee \big[C_E\sqrt{1 \vee [C_L + C_f + 1/2]}\big].$$

Let λ_0 *be defined as in* (3.2). *Then uniformly in the penalty level* λ *such that*

$$\lambda_0 \le \lambda \le C \cdot W_{\mathcal{U}}\sqrt{n\log p}, \tag{3.8}$$

we have that, for any $A > 1$ *with probability at least* $1 - \alpha - 4\gamma - 3p^{-A^2}$,

$$\sup_{u\in\mathcal{U}}\|J_u^{1/2}(\widehat{\beta}(u) - \beta(u))\| \le 8C \cdot \frac{(1+c_0)W_{\mathcal{U}}A}{\underline{f}^{1/2}\kappa_0}\cdot\sqrt{\frac{s\log(p\vee n)}{n}},$$

$$\sup_{u\in\mathcal{U}}\sqrt{\mathrm{E}_x[x'(\widehat{\beta}(u) - \beta(u))]^2} \le 8C \cdot \frac{(1+c_0)W_{\mathcal{U}}A}{\underline{f}\kappa_0}\cdot\sqrt{\frac{s\log(p\vee n)}{n}}$$

and

$$\sup_{u\in\mathcal{U}}\|\widehat{\beta}(u) - \beta(u)\| \le \frac{1 + c_0\sqrt{s/m}}{\kappa_m}\cdot 8C \cdot \frac{(1+c_0)W_{\mathcal{U}}A}{\underline{f}\kappa_0}\cdot\sqrt{\frac{s\log(p\vee n)}{n}},$$

provided s obeys the growth condition

$$2C \cdot (1 + c_0) W_{\mathcal{U}} A \cdot \sqrt{s \log(p \vee n)} < q \underline{f}^{1/2} \kappa_0 \sqrt{n}. \tag{3.9}$$

This result derives the rate of convergence of the ℓ_1-penalized quantile regression estimator in the intrinsic norm and other norms of interest uniformly in $u \in \mathcal{U}$ as well as uniformly in the penalty level λ in the range specified by (3.8), which includes our recommended choice of λ_0. We see that the rates of convergence for ℓ_1-QR generally depend on the number of significant regressors s, the logarithm of the number of regressors p, the strength of identification summarized by κ_0, κ_m, $\underline{f}$ and q, and the quantile indices of interest $\mathcal{U}$ (as expected, extreme quantiles can slow down the rates of convergence). These rate results parallel the results of [7] obtained for ℓ_1-penalized mean regression. Indeed, the role of the parameter $\underline{f}$ is similar to the role of the standard deviation of the disturbance in mean regression. It is worth noting, however, that our results do not rely on normality and homoscedasticity assumptions, and our proofs have to address the nonquadratic nature of the objective function, with parameter q controlling the quality of quadratization. This parameter q enters the results only through the growth restriction (3.9) on s. At this point, we refer the reader to Section 2.4 for a further discussion of this result in the context of the correlated normal design. Finally, we note that our proof combines the star-shaped geometry of the restricted set A_u with classical convexity arguments; this insight may be of interest in other problems.

PROOF OF THEOREM 2. We let

$$t := 8C \cdot \frac{(1 + c_0) W_{\mathcal{U}} A}{\underline{f}^{1/2} \kappa_0} \cdot \sqrt{\frac{s \log(p \vee n)}{n}}$$

and consider the following events:

(i) $\Omega_1 :=$ the event that (3.1) and $\widehat{\beta}(u) - \beta(u) \in A_u$, uniformly in $u \in \mathcal{U}$, hold;
(ii) $\Omega_2 :=$ the event that the bound on empirical error $\epsilon(t)$ in Lemma 5 holds;
(iii) $\Omega_3 :=$ the event in which $\Lambda(1 - \alpha | X) \leq \sqrt{1 + \log(1/\alpha)/\log p} \cdot C_\Lambda W_{\mathcal{U}} \times \sqrt{n \log p}$.

By the choice of λ and Lemma 3, $P(\Omega_1) \geq 1 - \alpha - \gamma$; by Lemma 5 $P(\Omega_2) \geq 1 - 3\gamma - 3p^{-A^2}$; and by Theorem 1 $P(\Omega_3) = 1$, hence $P(\bigcap_{k=1}^3 \Omega_k) \geq 1 - \alpha - 4\gamma - 3p^{-A^2}$.

Given the event $\bigcap_{k=1}^3 \Omega_k$, we want to show the event that

$$\exists u \in \mathcal{U} \qquad \| J_u^{1/2} (\widehat{\beta}(u) - \beta(u)) \| > t \tag{3.10}$$

is impossible, which will prove the first bound. The other two bounds then follow from Lemma 4 and the first bound. First, note that the event in (3.10) implies that

for some $u \in \mathcal{U}$

$$0 > \min_{\delta \in A_u, \|J_u^{1/2}\delta\| \geq t} \widehat{Q}_u(\beta(u)+\delta) - \widehat{Q}_u(\beta(u)) + \frac{\lambda\sqrt{u(1-u)}}{n}\big(\|\beta(u)+\delta\|_{1,n} - \|\beta(u)\|_{1,n}\big).$$

The key observation is that by convexity of $\widehat{Q}_u(\cdot) + \|\cdot\|_{1,n}\lambda\sqrt{u(1-u)}/n$ and by the fact that A_u is a cone, we can replace $\|J_u^{1/2}\delta\| \geq t$ by $\|J_u^{1/2}\delta\| = t$ in the above inequality and still preserve it:

$$0 > \min_{\delta \in A_u, \|J_u^{1/2}\delta\| = t} \widehat{Q}_u(\beta(u)+\delta) - \widehat{Q}_u(\beta(u)) + \frac{\lambda\sqrt{u(1-u)}}{n}\big(\|\beta(u)+\delta\|_{1,n} - \|\beta(u)\|_{1,n}\big).$$

Also, by inequality (3.4) in Lemma 4, for each $\delta \in A_u$

$$\|\beta(u)\|_{1,n} - \|\beta(u)+\delta\|_{1,n} \leq \|\delta_{T_u}\|_{1,n} \leq 2\|\delta_{T_u}\|_1 \leq 2\sqrt{s}\|J_u^{1/2}\delta\|/\underline{f}^{1/2}\kappa_0,$$

which then further implies

$$0 > \min_{\delta \in A_u, \|J_u^{1/2}\delta\| = t} \widehat{Q}_u(\beta(u)+\delta) - \widehat{Q}_u(\beta(u)) - \frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}}{\underline{f}^{1/2}\kappa_0}\|J_u^{1/2}\delta\|. \tag{3.11}$$

Also by Lemma 5, under our choice of $t \geq 1/[\underline{f}^{1/2}\kappa_0\sqrt{n}]$, $\log(L\underline{f}\kappa_0^2) \leq (C_L + C_f)\log(n \vee p)$, and under event Ω_2

$$\epsilon(t) \leq tC_E\sqrt{1 \vee [C_L + C_f + 1/2]}\frac{(1+c_0)A}{\underline{f}^{1/2}\kappa_0}\sqrt{\frac{s\log(p \vee n)}{n}}. \tag{3.12}$$

Therefore, we obtain from (3.11) and (3.12)

$$0 \geq \min_{\delta \in A_u, \|J_u^{1/2}\delta\| = t} Q_u(\beta(u)+\delta) - Q_u(\beta(u)) - \frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}}{\underline{f}^{1/2}\kappa_0}\|J_u^{1/2}\delta\| - tC_E\sqrt{1 \vee [C_L + C_f + 1/2]}\frac{(1+c_0)A}{\underline{f}^{1/2}\kappa_0}\sqrt{\frac{s\log(p \vee n)}{n}}.$$

Using the identifiability relation (3.7) stated in Lemma 4, we further get

$$0 > \frac{t^2}{4} \wedge (qt) - t\frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}}{\underline{f}^{1/2}\kappa_0} - tC_E\sqrt{1 \vee [C_L + C_f + 1/2]}\frac{(1+c_0)A}{\underline{f}^{1/2}\kappa_0}\sqrt{\frac{s\log(p \vee n)}{n}}.$$

Using the upper bound on λ under event Ω_3, we obtain

$$0 > \frac{t^2}{4} \wedge (qt) - tC\frac{2\sqrt{s\log p}}{\sqrt{n}}\frac{W_{\mathcal{U}}}{\underline{f}^{1/2}\kappa_0} - tC_E\sqrt{1 \vee [C_L + C_f + 1/2]}\frac{(1+c_0)A}{\underline{f}^{1/2}\kappa_0}\sqrt{\frac{s\log(p\vee n)}{n}}.$$

Note that qt cannot be smaller than $t^2/4$ under the growth condition (3.9) in the theorem. Thus, using also the lower bound on C given in the theorem, $W_{\mathcal{U}} \geq 1$, and $c_0 \geq 1$, we obtain the relation

$$0 > \frac{t^2}{4} - t \cdot 2C\frac{(1+c_0)W_{\mathcal{U}}A}{\underline{f}^{1/2}\kappa_0} \cdot \sqrt{\frac{s\log(p\vee n)}{n}} = 0,$$

which is impossible. □

3.3. *Sparsity properties*. Next, we derive sparsity properties of the solution to ℓ_1-penalized quantile regression. Fundamentally, sparsity is linked to the first order optimality conditions of (2.4) and therefore to the (sub)gradient of the criterion function. In the case of least squares, the gradient is a smooth (linear) function of the parameters. In the case of quantile regression, the gradient is a highly non-smooth (piece-wise constant) function. To control the sparsity of $\widehat{\beta}(u)$, we rely on empirical process arguments to approximate gradients by smooth functions. In particular, we crucially exploit the fact that the entropy of all m-dimensional submodels of the p-dimensional model is of order $m\log p$, which depends on p only logarithmically.

The statement of the results will depend on the maximal k-sparse eigenvalue of $\mathrm{E}[x_i x_i']$ and $\mathbb{E}_n[x_i x_i']$:

$$\varphi_{\max}(k) = \max_{\delta\neq 0, \|\delta\|_0 \leq k} \frac{\mathrm{E}[(x_i'\delta)^2]}{\delta'\delta} \quad \text{and}$$
$$\phi(k) = \sup_{\delta\neq 0, \|\delta\|_0 \leq k} \frac{\mathbb{E}_n[(x_i'\delta)^2]}{\delta'\delta} \vee \frac{\mathrm{E}[(x_i'\delta)^2]}{\delta'\delta}. \tag{3.13}$$

In order to establish our main sparsity result, we need two preliminary lemmas.

LEMMA 6 (Empirical pre-sparsity). *Let* $\widehat{s} = \sup_{u\in\mathcal{U}} \|\widehat{\beta}(u)\|_0$. *Under* D.1–D.4, *for any* $\lambda > 0$, *with probability at least* $1-\gamma$ *we have*

$$\widehat{s} \leq n \wedge p \wedge [4n^2\phi(\widehat{s})W_{\mathcal{U}}^2/\lambda^2].$$

In particular, if $\lambda \geq 2\sqrt{2}W_{\mathcal{U}}\sqrt{n\log(n\vee p)\phi(n/\log(n\vee p))}$ *then* $\widehat{s} \leq n/\log(n\vee p)$.

This lemma establishes an initial bound on the number of nonzero components $\widehat{s}$ as a function of λ and $\phi(\widehat{s})$. Restricting

$$\lambda \geq 2\sqrt{2}W_{\mathcal{U}}\sqrt{n\log(n\vee p)\phi\big(n/\log(n\vee p)\big)}$$

makes the term $\phi(n/\log(n\vee p))$ appear in subsequent bounds instead of the term $\phi(n)$, which in turn weakens some assumptions. Indeed, not only is the first term smaller than the second, but also there are designs of interest where the second term diverges while the first does not; for instance, in Design 1, if $p \geq 2n$, we have $\phi(n/\log(n\vee p)) \lesssim_P 1$ while $\phi(n) \gtrsim_P \sqrt{\log p}$ by [2].

The following lemma establishes a bound on the sparsity as a function of the rate of convergence.

LEMMA 7 (Empirical sparsity). *Assume* D.1–D.4 *and let*

$$r = \sup_{u\in\mathcal{U}} \big\| J_u^{1/2}\big(\widehat{\beta}(u) - \beta(u)\big)\big\|.$$

Then, for any $\varepsilon > 0$, *there is a constant* $K_\varepsilon \geq \sqrt{2}$ *such that with probability at least* $1-\varepsilon-\gamma$,

$$\frac{\sqrt{\widehat{s}}}{W_{\mathcal{U}}} \leq \mu(\widehat{s})\frac{n}{\lambda}(r\wedge 1) + \sqrt{\widehat{s}}K_\varepsilon\frac{\sqrt{n\log(n\vee p)\phi(\widehat{s})}}{\lambda},$$
$$\mu(k) := 2\sqrt{\varphi_{\max}(k)}(1\vee 2\bar{f}/\underline{f}^{1/2}).$$

Finally, we combine these results to establish the main sparsity result. In what follows, we define $\bar{\phi}_\varepsilon$ as a constant such that $\phi(n/\log(n\vee p)) \leq \bar{\phi}_\varepsilon$ with probability $1-\varepsilon$.

THEOREM 3 (Uniform sparsity bounds). *Let* $\varepsilon > 0$ *be any constant, assume* D.1–D.4 *hold, and let* λ *satisfy* $\lambda \geq \lambda_0$ *and*

$$K W_{\mathcal{U}}\sqrt{n\log(n\vee p)} \leq \lambda \leq K' W_{\mathcal{U}}\sqrt{n\log(n\vee p)}$$

for some constant $K' \geq K \geq 2K_\varepsilon\bar{\phi}_\varepsilon^{1/2}$, *for* K_ε *defined in Lemma* 7. *Then, for any* $A > 1$ *with probability at least* $1-\alpha-2\varepsilon-4\gamma-p^{-A^2}$,

$$\widehat{s} := \sup_{u\in\mathcal{U}} \|\widehat{\beta}(u)\|_0 \leq s\cdot[16\mu W_{\mathcal{U}}/\underline{f}^{1/2}\kappa_0]^2[(1+c_0)AK'/K]^2,$$

where $\mu := \mu(n/\log(n\vee p))$, *provided that* s *obeys the growth condition*

$$2K'(1+c_0)AW_{\mathcal{U}}\sqrt{s\log(n\vee p)} < q\,\underline{f}^{1/2}\kappa_0\sqrt{n}. \tag{3.14}$$

The theorem states that by setting the penalty level λ to be possibly higher than our initial recommended choice λ_0, we can control $\widehat{s}$, which will be crucial for good performance of the post-penalized estimator. As a corollary, we note that if (a) $\mu \lesssim 1$, (b) $1/(\underline{f}^{1/2}\kappa_0) \lesssim 1$ and (c) $\bar{\phi}_\varepsilon \lesssim 1$ for each $\varepsilon > 0$, then $\widehat{s} \lesssim s$ with a high probability, so the dimension of the selected model is about the same as the dimension of the true model. Conditions (a), (b) and (c) easily hold for the correlated normal design in Design 1. In particular, (c) follows from the concentration inequalities and from results in classical random matrix theory; see [2] for proofs. Therefore the possibly higher λ needed to achieve the stated sparsity bound does not slow down the rate of ℓ_1-QR in this case. The growth condition (3.14) on s is also weak in this case.

PROOF OF THEOREM 3. By the choice of K and Lemma 6, $\widehat{s} \leq n/\log(n \vee p)$ with probability $1 - \varepsilon$. With at least the same probability, the choice of λ yields

$$K_\varepsilon \frac{\sqrt{n \log(n \vee p)\phi(\widehat{s})}}{\lambda} \leq \frac{K_\varepsilon \bar{\phi}_\varepsilon^{1/2}}{K W_{\mathcal{U}}} \leq \frac{1}{2W_{\mathcal{U}}},$$

so that by virtue of Lemma 7 and by $\mu(\widehat{s}) \leq \mu := \mu(n/\log(n \vee p))$,

$$\frac{\sqrt{\widehat{s}}}{W_{\mathcal{U}}} \leq \mu \frac{(r \wedge 1)n}{\lambda} + \frac{\sqrt{\widehat{s}}}{2W_{\mathcal{U}}}$$

or

$$\frac{\sqrt{\widehat{s}}}{W_{\mathcal{U}}} \leq 2\mu \frac{(r \wedge 1)n}{\lambda},$$

with probability $1 - 2\varepsilon$. Since all conditions of Theorem 2 hold, we obtain the result by plugging in the upper bound on $r = \sup_{u \in \mathcal{U}} \| J_u^{1/2}(\widehat{\beta}(u) - \beta(u)) \|$ from Theorem 2. □

3.4. *Model selection properties*. Next, we turn to the model selection properties of ℓ_1-QR.

THEOREM 4 (Model selection properties of ℓ_1-QR). *Let* $r^o = \sup_{u \in \mathcal{U}} \| \widehat{\beta}(u) - \beta(u) \|$. *If* $\inf_{u \in \mathcal{U}} \min_{j \in T_u} |\beta_j(u)| > r^o$, *then*

$$T_u := \operatorname{support}(\beta(u)) \subseteq \widehat{T}_u := \operatorname{support}(\widehat{\beta}(u)) \qquad \text{for all } u \in \mathcal{U}. \tag{3.15}$$

Moreover, the hard-thresholded estimator $\bar{\beta}(u)$, *defined for any* $\gamma \geq 0$ *by*

$$\bar{\beta}_j(u) = \widehat{\beta}_j(u) 1\{|\widehat{\beta}_j(u)| > \gamma\}, \qquad u \in \mathcal{U},\ j = 1, \ldots, p, \tag{3.16}$$

provided that γ *is chosen such that* $r^o < \gamma < \inf_{u \in \mathcal{U}} \min_{j \in T_u} |\beta_j(u)| - r^o$, *satisfies*

$$\operatorname{support}(\bar{\beta}(u)) = T_u \qquad \text{for all } u \in \mathcal{U}.$$

These results parallel analogous results in [26] for mean regression. The first result says that if nonzero coefficients are well separated from zero, then the support of ℓ_1-QR includes the support of the true model. The inclusion of the true support in (3.15) is in general one-sided; the support of the estimator can include some unnecessary components having true coefficients equal to zero. The second result states that if the further conditions are satisfied, additional hard thresholding can eliminate inclusions of such unnecessary components. The value of the hard threshold must explicitly depend on the unknown value $\min_{j\in T_u}|\beta_j(u)|$, characterizing the separation of nonzero coefficients from zero. The additional conditions stated in this theorem are strong and perfect model selection appears quite unlikely in practice. Certainly it does not work in all real empirical examples we have explored. This motivates our analysis of the post-model-selected estimator under conditions that allow for imperfect model selection, including cases where we miss some nonzero components or have additional unnecessary components.

3.5. *The post-penalized estimator.* In this section, we establish a bound on the rate of convergence of the post-penalized estimator. The proof relies crucially on the identifiability and control of the empirical error over the sparse sets $\widetilde{A}_u(\widetilde{m}) := \{\delta \in \mathbb{R}^p : \|\delta_{T_u^c}\|_0 \le \widetilde{m}\}$.

LEMMA 8 (Sparse identifiability and control of empirical error). 1. *Suppose* D.1 *and* D.5 *hold. Then for all* $\delta \in \widetilde{A}_u(\widetilde{m})$, $u \in \mathcal{U}$, *and* $\widetilde{m} \le n$, *we have that*

$$Q_u(\beta(u)+\delta) - Q_u(\beta(u)) \ge \frac{\|J_u^{1/2}\delta\|^2}{4} \wedge (\widetilde{q}_{\widetilde{m}}\|J_u^{1/2}\delta\|). \tag{3.17}$$

2. *Suppose* D.1, D.2 *and* D.5 *hold and that* $|\bigcup_{u\in\mathcal{U}} T_u| \le n$. *Then for any* $\varepsilon > 0$, *there is a constant* C_ε *such that with probability at least* $1-\varepsilon$ *the empirical error*

$$\epsilon_u(\delta) := \big|\widehat{Q}_u(\beta(u)+\delta) - Q_u(\beta(u)+\delta) - \big(\widehat{Q}_u(\beta(u)) - Q_u(\beta(u))\big)\big|$$

obeys

$$\sup_{u\in\mathcal{U},\delta\in\widetilde{A}_u(\widetilde{m}),\delta\neq 0} \frac{\epsilon_u(\delta)}{\|\delta\|} \le C_\varepsilon\sqrt{\frac{(\widetilde{m}\log(n\vee p) + s\log n)\phi(\widetilde{m}+s)}{n}}$$

for all $\widetilde{m} \le n$.

In order to prove this lemma, we exploit the crucial fact that the entropy of all m-dimensional submodels of the p-dimensional model is of order $m\log p$, which depends on p only logarithmically. The following theorem establishes the properties of post-model-selection estimators.

THEOREM 5 (Uniform bounds on estimation error of post-ℓ_1-QR). *Assume the conditions of Theorem* 2 *hold, assume that* $|\bigcup_{u\in\mathcal{U}} T_u| \le n$, *and assume* D.5

holds with $\widehat{m} := \sup_{u\in\mathcal{U}} \|\widehat{\beta}_{T_u^c}(u)\|_0$ *with probability* $1-\varepsilon$. *Then for any* $\varepsilon > 0$ *there is a constant* C_ε *such that the bounds*

$$\begin{aligned}
&\sup_{u\in\mathcal{U}} \big\| J_u^{1/2}\big(\widetilde{\beta}(u) - \beta(u)\big)\big\| \\
&\qquad \le \frac{4C_\varepsilon\sqrt{\phi(\widehat{m}+s)}}{\underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}} \cdot \sqrt{\frac{\widehat{m}\log(n\vee p) + s\log n}{n}} \\
&\qquad\quad + \sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} \cdot \frac{4\sqrt{2(1+c_0)A}}{\underline{f}^{1/2}\kappa_0} \cdot C \cdot W_{\mathcal{U}} \sqrt{\frac{s\log(n\vee p)}{n}}, \qquad (3.18)\\
&\sup_{u\in\mathcal{U}} \sqrt{\mathrm{E}_x\big[x'\big(\widetilde{\beta}(u) - \beta(u)\big)\big]^2} \le \sup_{u\in\mathcal{U}} \big\| J_u^{1/2}\big(\widetilde{\beta}(u) - \beta(u)\big)\big\| / \underline{f}^{1/2}, \\
&\qquad \sup_{u\in\mathcal{U}} \|\widetilde{\beta}(u) - \beta(u)\| \le \sup_{u\in\mathcal{U}} \big\| J_u^{1/2}\big(\widetilde{\beta}(u) - \beta(u)\big)\big\| / \underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}},
\end{aligned}$$

hold with probability at least $1-\alpha-3\gamma-3p^{-A^2}-2\varepsilon$, *provided that* s *obeys the growth condition*

$$\begin{aligned}
&\widetilde{q}_{\widehat{m}} \frac{C_\varepsilon\sqrt{(\widehat{m}\log(n\vee p) + s\log n)\phi(\widehat{m}+s)}}{\sqrt{n}\,\underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}} \\
&\qquad + \sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} 2A(1+c_0) \cdot C^2 W_{\mathcal{U}}^2 \cdot \frac{s\log(p\vee n)}{n\underline{f}\kappa_0^2} \le \widetilde{q}_{\widehat{m}}^2.
\end{aligned}$$

This theorem describes the performance of post-ℓ_1-QR. However, an inspection of the proof reveals that it can be applied to any post-model selection estimator. From Theorem 5, we can conclude that in many interesting cases the rates of post-ℓ_1-QR could be the same or faster than the rate of ℓ_1-QR. Indeed, first consider the case where the model selection fails to contain the true model, that is, $\sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} = 1$ with a nonnegligible probability. If (a) $\widehat{m} \le \widehat{s} \lesssim_P s$, (b) $\phi(\widehat{m}+s) \lesssim_P 1$ and (c) the constants $\underline{f}$ and $\widetilde{\kappa}_{\widehat{m}}^2$ are of the same order as $\underline{f}$ and $\kappa_0\kappa_m$, respectively, then the rate of convergence of post-ℓ_1-QR is the same as the rate of convergence of ℓ_1-QR. Recall that Theorem 3 provides sufficient conditions needed to achieve (a), which hold in Design 1. Recall also that in Design 1, (b) holds by concentration of measure and classical results in random matrix theory, as shown in [2], and (c) holds by the calculations presented in Section 2. This verifies our claim regarding the performance of post-ℓ_1-QR in the overview, Section 2.4. The intuition for this result is that even though ℓ_1-QR misses true components, it does not miss very important ones, allowing post-ℓ_1-QR still to perform well. Second, consider the case where the model selection succeeds in containing the true model, that is, $\sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} = 0$ with probability approaching one, and that the number of unnecessary components obeys $\widehat{m} = o_P(s)$. In this case,

the rate of convergence of post-ℓ_1-QR can be faster than the rate of convergence of ℓ_1-QR. In the extreme case of perfect model selection, when $\widehat{m} = 0$ with a high probability, post-ℓ_1-QR becomes the oracle estimator with a high probability. We refer the reader to Section 2 for further discussion, and note that this result could be of interest in other problems.

PROOF OF THEOREM 5. Let

$$\widehat{\delta}(u) = \widehat{\beta}(u) - \beta(u), \qquad \widetilde{\delta}(u) := \widetilde{\beta}(u) - \beta(u), \qquad t_u := \|J_u^{1/2}\widetilde{\delta}(u)\|,$$

and B_n be a random variable such that $B_n = \sup_{u\in\mathcal{U}} \widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u))$. By the optimality of $\widehat{\beta}(u)$ in (3.16), with probability $1-\gamma$ we have uniformly in $u \in \mathcal{U}$

$$\begin{aligned}\widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u)) &\le \frac{\lambda\sqrt{u(1-u)}}{n}\big(\|\beta(u)\|_{1,n} - \|\widehat{\beta}(u)\|_{1,n}\big)\\ &\le \frac{\lambda\sqrt{u(1-u)}}{n}\|\widehat{\delta}_{T_u}(u)\|_{1,n} \qquad (3.19)\\ &\le \frac{\lambda\sqrt{u(1-u)}}{n}2\|\widehat{\delta}_{T_u}(u)\|_1,\end{aligned}$$

where the last term in (3.19) is bounded by

$$\begin{aligned}&\frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}\|J_u^{1/2}\widehat{\delta}(u)\|}{\underline{f}^{1/2}\kappa_0}\\ &\le \frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}}{\underline{f}^{1/2}\kappa_0}\sup_{u\in\mathcal{U}}\|J_u^{1/2}\big(\widehat{\beta}(u)-\beta(u)\big)\|,\end{aligned} \qquad (3.20)$$

using that $\|J_u^{1/2}\widehat{\delta}(u)\| \ge \underline{f}^{1/2}\kappa_0\|\widehat{\delta}_{T_u}(u)\|$ from (RE(c_0, 0)) implied by D.4. Therefore, by Theorem 2 we have

$$B_n \le \frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}}{\underline{f}^{1/2}\kappa_0}8C\cdot\frac{(1+c_0)W_{\mathcal{U}}A}{\underline{f}^{1/2}\kappa_0}\cdot\sqrt{\frac{s\log(p\vee n)}{n}}$$

with probability $1-\alpha-3\gamma-3p^{-A^2}$.

For every $u\in\mathcal{U}$, by optimality of $\widetilde{\beta}(u)$ in (2.5),

$$\begin{aligned}\widehat{Q}_u(\widetilde{\beta}(u)) - \widehat{Q}_u(\beta(u)) &\le 1\{T_u \not\subseteq \widehat{T}_u\}\big(\widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u))\big)\\ &\le 1\{T_u \not\subseteq \widehat{T}_u\}B_n.\end{aligned} \qquad (3.21)$$

Also, by Lemma 8, with probability at least $1-\varepsilon$, we have

$$\sup_{u\in\mathcal{U}}\frac{\epsilon_u(\widetilde{\delta}(u))}{\|\widetilde{\delta}(u)\|} \le C_\varepsilon\sqrt{\frac{(\widehat{m}\log(n\vee p) + s\log n)\phi(\widehat{m}+s)}{n}} =: A_{\varepsilon,n}. \qquad (3.22)$$

Recall that $\sup_{u\in\mathcal{U}} \|\widetilde{\delta}_{T_u^c}(u)\| \le \widehat{m} \le n$ so that by D.5 $t_u \ge \underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}\|\widetilde{\delta}(u)\|$ for all $u \in \mathcal{U}$ with probability $1-\varepsilon$. Thus, combining relations (3.21) and (3.22), for every $u \in \mathcal{U}$

$$Q_u(\widetilde{\beta}(u)) - Q_u(\beta(u)) \le t_u A_{\varepsilon,n}/[\underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}] + 1\{T_u \not\subseteq \widehat{T}_u\}B_n$$

with probability at least $1-2\varepsilon$. Invoking the sparse identifiability relation (3.17) of Lemma 8, with the same probability, for all $u \in \mathcal{U}$,

$$(t_u^2/4) \wedge (\widetilde{q}_{\widehat{m}} t_u) \le t_u A_{\varepsilon,n}/[\underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}] + 1\{T_u \not\subseteq \widehat{T}_u\}B_n.$$

We then conclude that under the assumed growth condition on s, this inequality implies

$$t_u \le 4A_{\varepsilon,n}/[\underline{f}^{1/2}\widetilde{\kappa}_{\widehat{m}}] + 1\{T_u \not\subseteq \widehat{T}_u\}\sqrt{4B_n \vee 0}$$

for every $u \in \mathcal{U}$, and the bounds stated in the theorem now follow from the definition of $\underline{f}$ and $\widetilde{\kappa}_m$. □

4. Empirical performance. In order to assess the finite sample practical performance of the proposed estimators, we conducted a Monte Carlo study. We will compare the performance of the ℓ_1-penalized, post-ℓ_1-penalized and the ideal oracle quantile regression estimators. Recall that the post-penalized estimator applies canonical quantile regression to the model selected by the penalized estimator. The oracle estimator applies canonical quantile regression to the true model. (Of course, such an estimator is not available outside Monte Carlo experiments.) We focus our attention on the model selection properties of the penalized estimator and biases and empirical risks of these estimators.

We begin by considering the following regression model:

$$y = x'\beta(0.5) + \varepsilon, \qquad \beta(0.5) = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \ldots, 0)',$$

where as in Design 1, $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Sigma)$, and the errors ε are independently and identically distributed $\varepsilon \sim N(0, \sigma^2)$. The dimension p of covariates x is 500, and the dimension s of the true model is 6, and the sample size n is 100. We set the regularization parameter λ equal to the 0.9-quantile of the pivotal random variable Λ, following our proposal in Section 2. The regressors are correlated with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. We consider two levels of noise, namely $\sigma = 1$ and $\sigma = 0.1$.

We summarize the model selection performance of the penalized estimator in Figures 1 and 2. In the left panels of the figures, we plot the frequencies of the dimensions of the selected model; in the right panels, we plot the frequencies of selecting the correct regressors. From the left panels, we see that the frequency of selecting a much larger model than the true model is very small in both designs. In the design with a larger noise, as the right panel of Figure 1 shows, the penalized quantile regression never selects the entire true model correctly, always

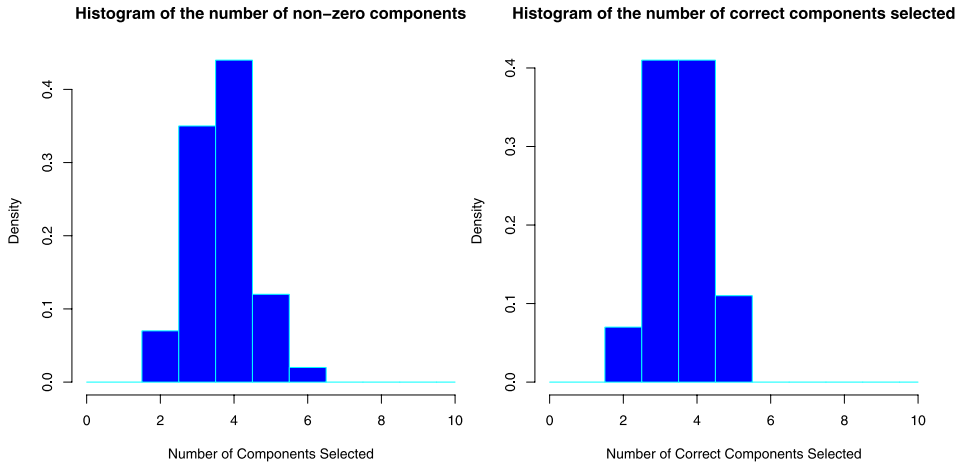

FIG. 1. *The figure summarizes the covariate selection results for the design with* $\sigma = 1$, *based on* 100 *Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible* 500 *covariates. The right panel plots the histogram for the number of significant covariates selected*; *there are in total* 6 *significant covariates amongst* 500 *covariates. The sample size for each repetition was* $n = 100$.

missing the regressors with small coefficients. However, it almost always includes the three regressors with the largest coefficients. (Notably, despite this partial fail-

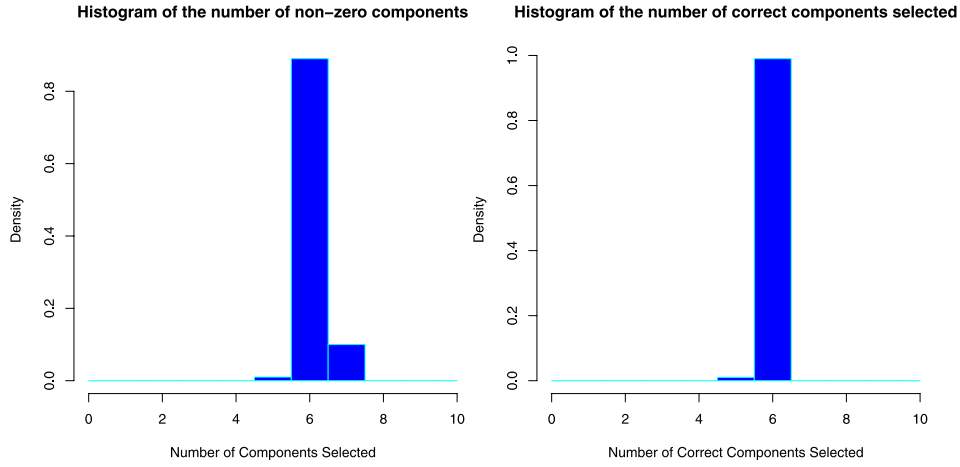

FIG. 2. *The figure summarizes the covariate selection results for the design with* $\sigma = 0.1$, *based on* 100 *Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible* 500 *covariates. The right panel plots the histogram for the number of significant covariates selected*; *there are in total* 6 *significant covariates amongst* 500 *covariates. The sample size for each repetition was* $n = 100$.

TABLE 1
Monte Carlo results. The table displays the average ℓ_0 and ℓ_1 norm of the estimators as well as mean bias and empirical risk. We obtained the results using 100 *Monte Carlo repetitions for each design*

	Mean ℓ_0-norm	**Mean ℓ_1-norm**	**Bias**	**Empirical risk**
		Design A ($\sigma = 1$)		
Penalized QR	3.67	1.28	0.92	1.22
Post-penalized QR	3.67	2.90	0.27	0.57
Oracle QR	6.00	3.31	0.03	0.33
		Design B ($\sigma = 0.1$)		
Penalized QR	6.09	2.98	0.13	0.19
Post-penalized QR	6.09	3.28	0.00	0.04
Oracle QR	6.00	3.28	0.00	0.03

ure of the model selection, post-penalized quantile regression still performs well, as we report below.) On the other hand, we see from the right panel of Figure 2 that in the design with a lower noise level penalized quantile regression rarely misses any component of the true support. These results confirm the theoretical results of Theorem 4, namely, that when the nonzero coefficients are well separated from zero, the penalized estimator should select a model that includes the true model as a subset. Moreover, these results also confirm the theoretical result of Theorem 3, namely, that the dimension of the selected model should be of the same stochastic order as the dimension of the true model. In summary, the model selection performance of the penalized estimator agrees very well with our theoretical results.

We summarize results on estimation performance in Table 1, which records for each estimator $\widetilde{\beta}$ the norm of the bias $\|\mathrm{E}[\widetilde{\beta}] - \beta_0\|$ and also the empirical risk $[\mathrm{E}[x_i'(\widetilde{\beta} - \beta_0)]^2]^{1/2}$ for recovering the regression function. Penalized quantile regression has a substantial bias, as we would expect from the definition of the estimator which penalizes large deviations of coefficients from zero. We see that the post-penalized quantile regression drastically improves upon the penalized quantile regression, particularly in terms of reducing the bias, which results in a much lower overall empirical risk. Notably, despite that under the higher noise level the penalized estimator never recovers the true model correctly the post-penalized estimator still performs well. This is because the penalized estimator always manages to select the most important regressors. We also see that the empirical risk of the post-penalized estimator is within a factor of $\sqrt{\log p}$ of the empirical risk of the oracle estimator, as we would expect from our theoretical results. Under the lower noise level, the post-penalized estimator performs almost identically to the ideal oracle estimator. We would expect this since in this case the penalized estimator selects the model especially well, making the post-penalized estimator nearly the oracle. In summary, we find the estimation performance of the penalized and post-penalized estimators to be in close agreement with our theoretical results.

APPENDIX A: PROOF OF THEOREM 1

PROOF OF THEOREM 1. We note $\Lambda \leq W_{\mathcal{U}} \max_{1\leq j\leq p} \sup_{u\in\mathcal{U}} n\mathbb{E}_n[(u - 1\{u_i \leq u\})x_{ij}/\widehat{\sigma}_j]$. For any $u \in \mathcal{U}$, $j \in \{1, \ldots, p\}$, we have by Lemma 1.5 in [24] that $P(|\mathbb{G}_n[(u - 1\{u_i \leq u\})x_{ij}/\widehat{\sigma}_j]| \geq \widetilde{K}) \leq 2\exp(-\widetilde{K}^2/2)$. Hence, by the symmetrization lemma for probabilities, Lemma 2.3.7 in [33], with $\widetilde{K} \geq 2\sqrt{\log 2}$ we have

$$
\begin{aligned}
&P\big(\Lambda > \widetilde{K}\sqrt{n}|X\big) \\
&\quad\leq 4P\Big(\sup_{u\in\mathcal{U}} \max_{1\leq j\leq p} |\mathbb{G}_n^o[(u - 1\{u_i \leq u\})x_{ij}/\widehat{\sigma}_j]| > \widetilde{K}/(4W_{\mathcal{U}})|X\Big) \qquad \text{(A.1)} \\
&\quad\leq 4p \max_{1\leq j\leq p} P\Big(\sup_{u\in\mathcal{U}} |\mathbb{G}_n^o[(u - 1\{u_i \leq u\})x_{ij}/\widehat{\sigma}_j]| > \widetilde{K}/(4W_{\mathcal{U}})|X\Big),
\end{aligned}
$$

where $\mathbb{G}_n^o$ denotes the symmetrized empirical process (see [33]) generated by the Rademacher variables ε_i, $i = 1, \ldots, n$, which are independent of $U = (u_1, \ldots, u_n)$ and $X = (x_1, \ldots, x_n)$. Let us condition on U and X, and define $\mathcal{F}_j = \{\varepsilon_i x_{ij}(u - 1\{u_i \leq u\})/\widehat{\sigma}_j : u \in \mathcal{U}\}$ for $j = 1, \ldots, p$. The VC dimension of $\mathcal{F}_j$ is at most 6. Therefore, by Theorem 2.6.7 of [33] for some universal constant $C_1' \geq 1$ the function class $\mathcal{F}_j$ with envelope function F_j obeys

$$
N(\varepsilon\|F_j\|_{\mathbb{P}_n,2}, \mathcal{F}_j, L_2(\mathbb{P}_n)) \leq n(\varepsilon, \mathcal{F}_j) = C_1' \cdot 6 \cdot (16e)^6(1/\varepsilon)^{10},
$$

where $N(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))$ denotes the minimal number of balls of radius ε with respect to the $L_2(\mathbb{P}_n)$ norm $\|\cdot\|_{\mathbb{P}_n,2}$ needed to cover the class of functions $\mathcal{F}$; see [33].

Conditional on the data $U = (u_1, \ldots, u_n)$ and $X = (x_1, \ldots, x_n)$, the symmetrized empirical process $\{\mathbb{G}_n^o(f), f \in \mathcal{F}_j\}$ is sub-Gaussian with respect to the $L_2(\mathbb{P}_n)$ norm by the Hoeffding inequality; see, for example, [33]. Since $\|F_j\|_{\mathbb{P}_n,2} \leq 1$ and $\rho(\mathcal{F}_j, \mathbb{P}_n) = \sup_{f\in\mathcal{F}_j} \|f\|_{\mathbb{P}_n,2}/\|F\|_{\mathbb{P}_n,2} \leq 1$, we have

$$
\begin{aligned}
&\|F_j\|_{\mathbb{P}_n,2} \int_0^{\rho(\mathcal{F}_j,\mathbb{P}_n)/4} \sqrt{\log n(\varepsilon, \mathcal{F}_j)}\, d\varepsilon \\
&\qquad\leq \bar{e} := (1/4)\sqrt{\log(6C_1'(16e)^6)} + (1/4)\sqrt{10\log 4}.
\end{aligned}
$$

By Lemma 16 with $D = 1$, there is a universal constant c such that for any $K \geq 1$:

$$
\begin{aligned}
&P\Big(\sup_{f\in\mathcal{F}_j} |\mathbb{G}_n^o(f)| > Kc\bar{e}|X, U\Big) \\
&\qquad\leq \int_0^{1/2} \varepsilon^{-1} n(\varepsilon, \mathcal{F}_j)^{-(K^2-1)}\, d\varepsilon \qquad \text{(A.2)} \\
&\qquad\leq (1/2)[6C_1'(16e)^6]^{-(K^2-1)} \frac{(1/2)^{10(K^2-1)}}{10(K^2-1)}.
\end{aligned}
$$

By (A.1) and (A.2) for any $k \geq 1$, we have

$$\begin{aligned} &P\big(\Lambda \geq k \cdot \big(4\sqrt{2}c\bar{e}\big) W_{\mathcal{U}} \sqrt{n \log p} | X\big) \\ &\quad \leq 4p \max_{1 \leq j \leq p} \mathrm{E}_U P\Big(\sup_{f \in \mathcal{F}_j} |\mathbb{G}_n^o(f)| > k\sqrt{2 \log p} c\bar{e} | X, U\Big) \\ &\quad \leq p^{-6k^2+1} \leq p^{-k^2+1} \end{aligned}$$

since $(2k^2 \log p - 1) \geq (\log 2 - 0.5)k^2 \log p$ for $p \geq 2$. Thus, result (i) holds with $C_\Lambda := 4\sqrt{2}c\bar{e}$. Result (ii) follows immediately by choosing

$$k = \sqrt{1 + \log(1/\alpha)/\log p}$$

to make the right-hand side of the display above equal to α. □

APPENDIX B: PROOFS OF LEMMAS 3–5 (USED IN THEOREM 2)

PROOF OF LEMMA 3 (Restricted set). 1. By condition D.3, with probability $1 - \gamma$, for every $j = 1, \ldots, p$ we have $1/2 \leq \widehat{\sigma}_j \leq 3/2$, which implies (3.1).

2. Denote the true rankscores by $a_i^*(u) = u - 1\{y_i \leq x_i'\beta(u)\}$ for $i = 1, \ldots, n$. Next, recall that $\widehat{Q}_u(\cdot)$ is a convex function and $\mathbb{E}_n[x_i a_i^*(u)] \in \partial \widehat{Q}_u(\beta(u))$. Therefore, we have

$$\widehat{Q}_u(\widehat{\beta}(u)) \geq \widehat{Q}_u(\beta(u)) + \mathbb{E}_n[x_i a_i^*(u)]'\big(\widehat{\beta}(u) - \beta(u)\big).$$

Let $\widehat{D} = \mathrm{diag}[\widehat{\sigma}_1, \ldots, \widehat{\sigma}_p]$ and note that $\lambda\sqrt{u(1-u)}(c_0 - 3)/(c_0 + 3) \geq n\|\widehat{D}^{-1} \times \mathbb{E}_n[x_i a_i^*(u)]\|_\infty$ with probability at least $1 - \alpha$. By optimality of $\widehat{\beta}(u)$ for the ℓ_1-penalized problem, we have

$$\begin{aligned} 0 &\leq \widehat{Q}_u(\beta(u)) - \widehat{Q}_u(\widehat{\beta}(u)) \\ &\quad + \frac{\lambda\sqrt{u(1-u)}}{n}\|\beta(u)\|_{1,n} - \frac{\lambda\sqrt{u(1-u)}}{n}\|\widehat{\beta}(u)\|_{1,n} \\ &\leq \big|\mathbb{E}_n[x_i a_i^*(u)]'\big(\widehat{\beta}(u) - \beta(u)\big)\big| \\ &\quad + \frac{\lambda\sqrt{u(1-u)}}{n}\big(\|\beta(u)\|_{1,n} - \|\widehat{\beta}(u)\|_{1,n}\big) \\ &= \|\widehat{D}^{-1}\mathbb{E}_n[x_i a_i^*(u)]\|_\infty \big\|\widehat{D}\big(\widehat{\beta}(u) - \beta(u)\big)\big\|_1 \\ &\quad + \frac{\lambda\sqrt{u(1-u)}}{n}\big(\|\beta(u)\|_{1,n} - \|\widehat{\beta}(u)\|_{1,n}\big) \\ &\leq \frac{\lambda\sqrt{u(1-u)}}{n}\sum_{j=1}^{p}\Big(\frac{c_0 - 3}{c_0 + 3}\widehat{\sigma}_j|\widehat{\beta}_j(u) - \beta_j(u)| + \widehat{\sigma}_j|\beta_j(u)| - \widehat{\sigma}_j|\widehat{\beta}_j(u)|\Big), \end{aligned}$$

with probability at least $1-\alpha$. After canceling $\lambda\sqrt{u(1-u)}/n$, we obtain

$$\begin{aligned}\left(1-\frac{c_0-3}{c_0+3}\right)\|\widehat{\beta}(u)-\beta(u)\|_{1,n}\\ \leq \sum_{j=1}^{p}\widehat{\sigma}_j\big(|\widehat{\beta}_j(u)-\beta_j(u)|+|\beta_j(u)|-|\widehat{\beta}_j(u)|\big).\end{aligned}\tag{B.1}$$

Furthermore, since $|\widehat{\beta}_j(u)-\beta_j(u)|+|\beta_j(u)|-|\widehat{\beta}_j(u)|=0$ if $\beta_j(u)=0$, that is, $j\in T_u^c$,

$$\sum_{j=1}^{p}\widehat{\sigma}_j\big(|\widehat{\beta}_j(u)-\beta_j(u)|+|\beta_j(u)|-|\widehat{\beta}_j(u)|\big)\leq 2\|\widehat{\beta}_{T_u}(u)-\beta(u)\|_{1,n}.\tag{B.2}$$

Equations (B.1) and (B.2) establish that $\|\widehat{\beta}_{T_u^c}(u)\|_{1,n}\leq(c_0/3)\|\widehat{\beta}_{T_u}(u)-\beta(u)\|_{1,n}$ with probability at least $1-\alpha$. In turn, by part 1 of this lemma, $\|\widehat{\beta}_{T_u^c}(u)\|_{1,n}\geq(1/2)\|\widehat{\beta}_{T_u^c}(u)\|_1$ and $\|\widehat{\beta}_{T_u}(u)-\beta(u)\|_{1,n}\leq(3/2)\|\widehat{\beta}_{T_u}(u)-\beta(u)\|_1$, which holds with probability at least $1-\gamma$. Intersection of these two event holds with probability at least $1-\alpha-\gamma$. Finally, by Lemma 9, $\|\widehat{\beta}(u)\|_0\leq n$ with probability 1 uniformly in $u\in\mathcal{U}$. □

PROOF OF LEMMA 4 (Identification in population).

1. PROOFS OF CLAIMS (3.3)–(3.5). By (RE(c_0,m)) and by $\delta\in A_u$

$$\begin{aligned}\|J_u^{1/2}\delta\|&\geq\|(\mathrm{E}[x_ix_i'])^{1/2}\delta\|\underline{f}^{1/2}\geq\|\delta_{T_u}\|\underline{f}^{1/2}\kappa_0\geq\frac{\underline{f}^{1/2}\kappa_0}{\sqrt{s}}\|\delta_{T_u}\|_1\\ &\geq\frac{\underline{f}^{1/2}\kappa_0}{\sqrt{s}(1+c_0)}\|\delta\|_1.\end{aligned}$$

2. PROOF OF CLAIM (3.6). Proceeding similarly to [7], we note that the kth largest in absolute value component of $\delta_{T_u^c}$ is less than $\|\delta_{T_u^c}\|_1/k$. Therefore, by $\delta\in A_u$ and $|T_u|\leq s$

$$\begin{aligned}\|\delta_{(T_u\cup\overline{T}_u(\delta,m))^c}\|^2&\leq\sum_{k\geq m+1}\frac{\|\delta_{T_u^c}\|_1^2}{k^2}\leq\frac{\|\delta_{T_u^c}\|_1^2}{m}\leq c_0^2\frac{\|\delta_{T_u}\|_1^2}{m}\\ &\leq c_0^2\|\delta_{T_u}\|^2\frac{s}{m}\leq c_0^2\|\delta_{T_u\cup\overline{T}_u(\delta,m)}\|^2\frac{s}{m},\end{aligned}$$

so that $\|\delta\|\leq(1+c_0\sqrt{s/m})\|\delta_{T_u\cup\overline{T}_u(\delta,m)}\|$; and the last term is bounded by (RE(c_0,m)),

$$\big(1+c_0\sqrt{s/m}\big)\|(\mathrm{E}[x_ix_i'])^{1/2}\delta\|/\kappa_m\leq\big(1+c_0\sqrt{s/m}\big)\|J_u^{1/2}\delta\|/[\underline{f}^{1/2}\kappa_m].$$

3. PROOF OF CLAIM (3.7) proceeds in two steps. Step 1 (Minoration). Define the maximal radius over which the criterion function can be minorated by a quadratic function

$$r_{A_u} = \sup_r \Big\{ r : Q_u\big(\beta(u) + \widetilde{\delta}\big) - Q_u(\beta(u)) \geq \frac{1}{4}\|J_u^{1/2}\widetilde{\delta}\|^2 \quad \text{for all } \widetilde{\delta} \in A_u, \|J_u^{1/2}\widetilde{\delta}\| \leq r \Big\}.$$

Step 2 below shows that $r_{A_u} \geq 4q$. By construction of r_{A_u} and the convexity of Q_u,

$$\begin{aligned} &Q_u\big(\beta(u)+\delta\big) - Q_u(\beta(u)) \\ &\quad \geq \frac{\|J_u^{1/2}\delta\|^2}{4} \wedge \left\{ \frac{\|J_u^{1/2}\delta\|}{r_{A_u}} \cdot \inf_{\widetilde{\delta}\in A_u, \|J_u^{1/2}\widetilde{\delta}\| \geq r_{A_u}} Q_u\big(\beta(u)+\widetilde{\delta}\big) - Q_u(\beta(u)) \right\} \\ &\quad \geq \frac{\|J_u^{1/2}\delta\|^2}{4} \wedge \left\{ \frac{\|J_u^{1/2}\delta\|}{r_{A_u}} \frac{r_{A_u}^2}{4} \right\} \\ &\quad \geq \frac{\|J_u^{1/2}\delta\|^2}{4} \wedge \{q\|J_u^{1/2}\delta\|\} \qquad \text{for any } \delta \in A_u. \end{aligned}$$

Step 2. ($r_{A_u} \geq 4q$). Let $F_{y|x}$ denote the conditional distribution of y given x. From [17], for any two scalars w and v we have that

$$\begin{aligned} \rho_u(w-v) - \rho_u(w) &= -v(u - 1\{w \leq 0\}) \\ &\quad + \int_0^v (1\{w \leq z\} - 1\{w \leq 0\})\, dz. \end{aligned} \tag{B.3}$$

Using (B.3) with $w = y - x'\beta(u)$ and $v = x'\delta$, we conclude $\mathrm{E}[-v(u - 1\{w \leq 0\})] = 0$. Using the law of iterated expectations and mean value expansion, we obtain for $\widetilde{z}_{x,z} \in [0, z]$

$$\begin{aligned} &Q_u\big(\beta(u)+\delta\big) - Q_u(\beta(u)) \\ &\quad = \mathrm{E}\left[\int_0^{x'\delta} F_{y|x}\big(x'\beta(u) + z\big) - F_{y|x}(x'\beta(u))\, dz \right] \\ &\quad = \mathrm{E}\left[\int_0^{x'\delta} z f_{y|x}(x'\beta(u)) + \frac{z^2}{2} f'_{y|x}\big(x'\beta(u) + \widetilde{z}_{x,z}\big)\, dz \right] \\ &\quad \geq \frac{1}{2}\|J_u^{1/2}\delta\|^2 - \frac{1}{6}\bar{f}'\mathrm{E}[|x'\delta|^3] \\ &\quad \geq \frac{1}{4}\|J_u^{1/2}\delta\|^2 + \frac{1}{4}\underline{f}\mathrm{E}[|x'\delta|^2] - \frac{1}{6}\bar{f}'\mathrm{E}[|x'\delta|^3]. \end{aligned} \tag{B.4}$$

Note that for $\delta \in A_u$, if $\|J_u^{1/2}\delta\| \leq 4q \leq (3/2)\cdot(\underline{f}^{3/2}/\bar{f}')\cdot \inf_{\delta\in A_u,\delta\neq 0} \mathrm{E}[|x'\delta|^2]^{3/2}/\mathrm{E}[|x'\delta|^3]$, it follows that $(1/6)\,\bar{f}'\mathrm{E}[|x'\delta|^3] \leq (1/4)\,\underline{f}\,\mathrm{E}[|x'\delta|^2]$. This and (B.4) imply $r_{A_u} \geq 4q$. □

PROOF OF LEMMA 5 (Control of empirical error). We divide the proof in four steps.

Step 1 (Main argument). Let

$$\begin{aligned}\mathcal{A}(t) &:= \epsilon(t)\sqrt{n}\\ &= \sup_{u\in\mathcal{U},\,\|J_u^{1/2}\delta\|\leq t,\delta\in A_u} \bigl|\mathbb{G}_n\bigl[\rho_u\bigl(y_i - x_i'(\beta(u)+\delta)\bigr) - \rho_u\bigl(y_i - x_i'\beta(u)\bigr)\bigr]\bigr|.\end{aligned}$$

Let Ω_1 be the event in which $\max_{1\leq j\leq p}|\widehat{\sigma}_j - 1| \leq 1/2$, where $P(\Omega_1) \geq 1-\gamma$.

In order to apply the symmetrization lemma, Lemma 2.3.7 in [33], to bound the tail probability of $\mathcal{A}(t)$ first note that for any fixed $\delta \in A_u$, $u \in \mathcal{U}$ we have

$$\mathrm{var}\bigl(\mathbb{G}_n\bigl[\rho_u\bigl(y_i - x_i'(\beta(u)+\delta)\bigr) - \rho_u\bigl(y_i - x_i'\beta(u)\bigr)\bigr]\bigr) \leq \mathrm{E}[(x_i'\delta)^2] \leq t^2/\underline{f}.$$

Then application of the symmetrization lemma for probabilities, Lemma 2.3.7 in [33], yields

$$\begin{aligned}P\bigl(\mathcal{A}(t) \geq M\bigr) &\leq \frac{2P(\mathcal{A}^o(t) \geq M/4)}{1 - t^2/(\underline{f}M^2)}\\ &\leq \frac{2P(\mathcal{A}^o(t) \geq M/4|\Omega_1) + 2P(\Omega_1^c)}{1 - t^2/(\underline{f}M^2)},\end{aligned} \tag{B.5}$$

where $\mathcal{A}^o(t)$ is the symmetrized version of $\mathcal{A}(t)$, constructed by replacing the empirical process $\mathbb{G}_n$ with its symmetrized version $\mathbb{G}_n^o$, and $P(\Omega_1^c) \leq \gamma$. We set $M > M_1 := t(3/\underline{f})^{1/2}$, which makes the denominator on right-hand side of (B.5) greater than $2/3$. Further, Step 3 below shows that $P(\mathcal{A}^o(t) \geq M/4|\Omega_1) \leq p^{-A^2}$ for

$$\begin{aligned}M/4 \geq M_2 &:= t\cdot A\cdot 18\sqrt{2}\cdot\Gamma\cdot\sqrt{2\log p + \log\bigl(2 + 4\sqrt{2}L\underline{f}^{1/2}\kappa_0/t\bigr)},\\ \Gamma &= \sqrt{s}(1+c_0)/[\underline{f}^{1/2}\kappa_0].\end{aligned}$$

We conclude that with probability at least $1 - 3\gamma - 3p^{-A^2}$, $\mathcal{A}(t) \leq M_1 \vee (4M_2)$.

Therefore, there is a universal constant C_E such that with probability at least $1 - 3\gamma - 3p^{-A^2}$,

$$\mathcal{A}(t) \leq t\cdot C_E\cdot\frac{(1+c_0)A}{\underline{f}^{1/2}\kappa_0}\sqrt{s\log(p \vee [L\underline{f}^{1/2}\kappa_0/t])}$$

and the result follows.

Step 2 [Bound on $P(\mathcal{A}^o(t) \geq K|\Omega_1)$]. We begin by noting that Lemmas 3 and 4 imply that $\|\delta\|_{1,n} \leq \frac{3}{2}\sqrt{s}(1+c_0)\|J_u^{1/2}\delta\|/[\underline{f}^{1/2}\kappa_0]$ so that for all $u \in \mathcal{U}$

$$\begin{aligned} \{\delta \in A_u : \|J_u^{1/2}\delta\| \leq t\} &\subseteq \{\delta \in \mathbb{R}^p : \|\delta\|_{1,n} \leq 2t\Gamma\}, \\ \Gamma &:= \sqrt{s}(1+c_0)/[\underline{f}^{1/2}\kappa_0]. \end{aligned} \tag{B.6}$$

Further, we let $\mathcal{U}_k = \{\widehat{u}_1, \ldots, \widehat{u}_k\}$ be an ε-net of quantile indices in $\mathcal{U}$ with

$$\varepsilon \leq t\Gamma/(2\sqrt{2s}L) \quad \text{and} \quad k \leq 1/\varepsilon. \tag{B.7}$$

By $\rho_u(y_i - x_i'(\beta(u)+\delta)) - \rho_u(y_i - x_i'\beta(u)) = ux_i'\delta + w_i(x_i'\delta, u)$, for $w_i(b,u) := (y_i - x_i'\beta(u) - b)_- - (y_i - x_i'\beta(u))_-$, and by (B.6) we have that $\mathcal{A}^o(t) \leq \mathcal{B}^o(t) + \mathcal{C}^o(t)$, where

$$\mathcal{B}^o(t) := \sup_{u\in\mathcal{U}, \|\delta\|_{1,n}\leq 2t\Gamma} |\mathbb{G}_n^o[x_i'\delta]| \quad \text{and} \quad \mathcal{C}^o(t) := \sup_{u\in\mathcal{U}, \|\delta\|_{1,n}\leq 2t\Gamma} |\mathbb{G}_n^o[w_i(\delta,u)]|.$$

Then we compute the bounds

$$\begin{aligned} &P[\mathcal{B}^o(t) > K|\Omega_1] \\ &\quad \leq \min_{\lambda\geq 0} e^{-\lambda K}\mathrm{E}\big[e^{\lambda\mathcal{B}^o(t)}|\Omega_1\big] \qquad \text{by Markov} \\ &\quad \leq \min_{\lambda\geq 0} e^{-\lambda K} 2p\exp\big((2\lambda t\Gamma)^2/2\big) \qquad \text{by Step 3} \\ &\quad \leq 2p\exp\big(-K^2/(2\sqrt{2}t\Gamma)^2\big) \qquad \text{by setting } \lambda = K/(2t\Gamma)^2 \\ &P[\mathcal{C}^o(t) > K|\Omega_1] \\ &\quad \leq \min_{\lambda\geq 0} e^{-\lambda K}\mathrm{E}\big[e^{\lambda\mathcal{C}^o(t)}|\Omega_1, X\big] \qquad \text{by Markov} \\ &\quad \leq \min_{\lambda\geq 0} \exp(-\lambda K)2(p/\varepsilon)\exp\big((16\lambda t\Gamma)^2/2\big) \qquad \text{by Step 4} \\ &\quad \leq \varepsilon^{-1}2p\exp\big(-K^2/(16\sqrt{2}t\Gamma)^2\big) \qquad \text{by setting } \lambda = K/(16t\Gamma)^2, \end{aligned}$$

so that

$$\begin{aligned} &P\big[\mathcal{A}^o(t) > 2\sqrt{2}K + 16\sqrt{2}K|\Omega_1\big] \\ &\quad \leq P\big[\mathcal{B}^o(t) > 2\sqrt{2}K|\Omega_1\big] + P\big[\mathcal{C}^o(t) > 16\sqrt{2}K|\Omega_1\big] \\ &\quad \leq 2p(1+\varepsilon^{-1})\exp\big(-K^2/(t\Gamma)^2\big). \end{aligned}$$

Setting $K = A\cdot t\cdot\Gamma\cdot\sqrt{\log\{2p^2(1+\varepsilon^{-1})\}}$, for $A \geq 1$, we get $P[\mathcal{A}^o(t) \geq 18\sqrt{2}\times K|\Omega_1] \leq p^{-A^2}$.

Step 3 (Bound on $\mathrm{E}[e^{\lambda \mathcal{B}^o(t)}|\Omega_1]$). We bound

$$\begin{aligned}\mathrm{E}\big[e^{\lambda \mathcal{B}^o(t)}|\Omega_1\big] &\le \mathrm{E}\Big[\exp\Big(2\lambda t\Gamma \max_{j\le p}|\mathbb{G}_n^o(x_{ij})/\widehat{\sigma}_j|\Big)|\Omega_1\Big] \\ &\le 2p \max_{j\le p} \mathrm{E}\big[\exp\big(2\lambda t\Gamma \mathbb{G}_n^o(x_{ij})/\widehat{\sigma}_j\big)|\Omega_1\big] \\ &\le 2p \exp\big((2\lambda t\Gamma)^2/2\big),\end{aligned}$$

where the first inequality follows from $|\mathbb{G}_n^o[x_i'\delta]| \le 2\|\delta\|_{1,n} \max_{1\le j\le p}|\mathbb{G}_n^o(x_{ij})/\widehat{\sigma}_j|$ holding under event Ω_1, the penultimate inequality follows from the simple bound

$$\mathrm{E}\big[\max_{j\le p} e^{|z_j|}\big] \le p \max_{j\le p} \mathrm{E}\big[e^{|z_j|}\big] \le p \max_{j\le p} \mathrm{E}[e^{z_j} + e^{-z_j}] \le 2p \max_{j\le p} \mathrm{E}[e^{z_j}]$$

holding for symmetric random variables z_j, and the last inequality follows from the law of iterated expectations and from $\mathrm{E}[\exp(2\lambda t\Gamma \mathbb{G}_n^o(x_{ij})/\widehat{\sigma}_j)|\Omega_1, X] \le \exp((2\lambda t\Gamma)^2/2)$ holding by the Hoeffding inequality (more precisely, by the intermediate step in the proof of the Hoeffding inequality; see, e.g., page 100 in [33]). Here, $\mathrm{E}[\cdot|\Omega_1, X]$ denotes the expectation over the symmetrizing Rademacher variables entering the definition of the symmetrized process $\mathbb{G}_n^o$.

Step 4 (Bound on $\mathrm{E}[e^{\lambda \mathcal{C}^o(t)}|\Omega_1]$). We bound

$$\begin{aligned}\mathcal{C}^o(t) &\le \sup_{u\in\mathcal{U}, |u-\widehat{u}|\le\varepsilon, \widehat{u}\in\mathcal{U}_k} \sup_{\|\delta\|_{1,n}\le 2t\Gamma} \big|\mathbb{G}_n^o\big[w_i\big(x_i'\big(\delta+\beta(u)-\beta(\widehat{u})\big),\widehat{u}\big)\big]\big| \\ &\quad + \sup_{u\in\mathcal{U}, |u-\widehat{u}|\le\varepsilon, \widehat{u}\in\mathcal{U}_k} \big|\mathbb{G}_n\big[w_i\big(x_i'\big(\beta(u)-\beta(\widehat{u})\big),\widehat{u}\big)\big]\big| \\ &\le 2 \sup_{\widehat{u}\in\mathcal{U}_k, \|\delta\|_{1,n}\le 4t\Gamma} |\mathbb{G}_n^o[w_i(x_i'\delta,\widehat{u})]| =: \mathcal{D}^o(t),\end{aligned}$$

where the first inequality is elementary, and the second inequality follows from the inequality

$$\sup_{|u-\widehat{u}|\le\varepsilon} \|\beta(u)-\beta(\widehat{u})\|_{1,n} \le \sqrt{2s}L\Big(2\max_{1\le j\le p}\sigma_j\Big)\varepsilon \le \sqrt{2s}L(2\cdot 3/2)\varepsilon \le 2t\Gamma,$$

holding by our choice (B.7) of ε and by event Ω_1.

Next, we bound $\mathrm{E}[e^{\lambda \mathcal{D}^o(t)}|\Omega_1]$

$$\begin{aligned}\mathrm{E}\big[e^{\lambda \mathcal{D}^o(t)}|\Omega_1\big] &\le (1/\varepsilon) \max_{\widehat{u}\in\mathcal{U}_k} \mathrm{E}\Big[\exp\Big(2\lambda \sup_{\|\delta\|_{1,n}\le 4t\Gamma} |\mathbb{G}_n^o[w_i(x_i'\delta,\widehat{u})]|\Big)|\Omega_1\Big] \\ &\le (1/\varepsilon) \max_{\widehat{u}\in\mathcal{U}_k} \mathrm{E}\Big[\exp\Big(4\lambda \sup_{\|\delta\|_{1,n}\le 4t\Gamma} |\mathbb{G}_n^o[x_i'\delta]|\Big)|\Omega_1\Big] \\ &\le 2(p/\varepsilon) \max_{j\le p} \mathrm{E}\big[\exp\big(16\lambda t\Gamma \mathbb{G}_n^o(x_{ij})/\widehat{\sigma}_j\big)|\Omega_1\big] \\ &\le 2(p/\varepsilon) \exp\big((16\lambda t\Gamma)^2/2\big),\end{aligned}$$

where the first inequality follows from the definition of w_i and by $k \leq 1/\varepsilon$, the second inequality follows from the exponential moment inequality for contractions (Theorem 4.12 of Ledoux and Talagrand [24]) and from the contractive property $|w_i(a, \widehat{u}) - w_i(b, \widehat{u})| \leq |a - b|$, and the last two inequalities follow exactly as in Step 3. □

APPENDIX C: PROOFS OF LEMMAS 6, 7 (USED IN THEOREM 3)

In order to characterize the sparsity properties of $\widehat{\beta}(u)$, we will exploit the fact that (2.4) can be written as the following linear programming problem:

$$\begin{aligned} &\min_{\xi^+,\xi^-,\beta^+,\beta^- \in \mathbb{R}_+^{2n+2p}} \mathbb{E}_n[u\xi_i^+ + (1-u)\xi_i^-] \\ &\quad + \frac{\lambda\sqrt{u(1-u)}}{n} \sum_{j=1}^{p} \widehat{\sigma}_j(\beta_j^+ + \beta_j^-) \\ &\xi_i^+ - \xi_i^- = y_i - x_i'(\beta^+ - \beta^-), \qquad i = 1, \ldots, n. \end{aligned} \tag{C.1}$$

Our theoretical analysis of the sparsity of $\widehat{\beta}(u)$ relies on the dual of (C.1):

$$\begin{aligned} &\max_{a \in \mathbb{R}^n} \mathbb{E}_n[y_i a_i] | \mathbb{E}_n[x_{ij} a_i] | \leq \lambda\sqrt{u(1-u)}\widehat{\sigma}_j/n, \qquad j = 1, \ldots, p, \\ &\qquad (u-1) \leq a_i \leq u, \qquad i = 1, \ldots, n. \end{aligned} \tag{C.2}$$

The dual program maximizes the correlation between the response variable and the rank scores subject to the condition requiring the rank scores to be approximately uncorrelated with the regressors. The optimal solution $\widehat{a}(u)$ to (C.2) plays a key role in determining the sparsity of $\widehat{\beta}(u)$.

LEMMA 9 (Signs and interpolation property). (1) *For any* $j \in \{1, \ldots, p\}$

$$\begin{aligned} &\widehat{\beta}_j(u) > 0 \quad \textit{iff} \quad \mathbb{E}_n[x_{ij}\widehat{a}_i(u)] = \lambda\sqrt{u(1-u)}\widehat{\sigma}_j/n, \\ &\widehat{\beta}_j(u) < 0 \quad \textit{iff} \quad \mathbb{E}_n[x_{ij}\widehat{a}_i(u)] = -\lambda\sqrt{u(1-u)}\widehat{\sigma}_j/n. \end{aligned} \tag{C.3}$$

(2) $\|\widehat{\beta}(u)\|_0 \leq n \wedge p$ *uniformly over* $u \in \mathcal{U}$. (3) *If* $y_1, \ldots, y_n$ *are absolutely continuous conditional on* $x_1, \ldots, x_n$, *then the number of interpolated data points,* $I_u = |\{i : y_i = x_i'\widehat{\beta}(u)\}|$, *is equal to* $\|\widehat{\beta}(u)\|_0$ *with probability one uniformly over* $u \in \mathcal{U}$.

PROOF OF LEMMA 9. Step 1. Part (1) follows from the complementary slackness condition for linear programming problems; see Theorem 4.5 of [6]. Step 2. For proof of part (2), see [2]. □

PROOF OF LEMMA 6 (Empirical pre-sparsity). That $\widehat{s} \leq n \wedge p$ follows from Lemma 9. We proceed to show the last bound.

Let $\widehat{a}(u)$ be the solution of the dual problem (C.2), $\widehat{T}_u = \text{support}(\widehat{\beta}(u))$, and $\widehat{s}_u = \|\widehat{\beta}(u)\|_0 = |\widehat{T}_u|$. For any $j \in \widehat{T}_u$, from (C.3) we have $(X'\widehat{a}(u))_j = \text{sign}(\widehat{\beta}_j(u))\lambda\widehat{\sigma}_j\sqrt{u(1-u)}$ and, for $j \notin \widehat{T}_u$ we have $\text{sign}(\widehat{\beta}_j(u)) = 0$. Therefore, by the Cauchy–Schwarz inequality, and by D.3, with probability $1-\gamma$ we have

$$\begin{aligned}\widehat{s}_u\lambda &= \text{sign}(\widehat{\beta}(u))'\,\text{sign}(\widehat{\beta}(u))\lambda \le \text{sign}(\widehat{\beta}(u))'(X'\widehat{a}(u))\Big/\min_{j=1,\dots,p}\widehat{\sigma}_j\sqrt{u(1-u)}\\ &\le 2\|X\,\text{sign}(\widehat{\beta}(u))\|\|\widehat{a}(u)\|/\sqrt{u(1-u)}\\ &\le 2\sqrt{n\phi(\widehat{s}_u)}\|\text{sign}(\widehat{\beta}(u))\|\|\widehat{a}(u)\|/\sqrt{u(1-u)},\end{aligned}$$

where we used that $\|\text{sign}(\widehat{\beta}(u))\|_0 = \widehat{s}_u$ and $\min_{1\le j\le p}\widehat{\sigma}_j \ge 1/2$ with probability $1-\gamma$. Since $\|\widehat{a}(u)\| \le \sqrt{n}$, and $\|\text{sign}(\widehat{\beta}(u))\| = \sqrt{\widehat{s}_u}$ we have $\widehat{s}_u\lambda \le 2n\sqrt{\widehat{s}_u\phi(\widehat{s}_u)}W_{\mathcal{U}}$. Taking the supremum over $u \in \mathcal{U}$ on both sides yields the first result.

To establish the second result, note that $\widehat{s} \le \bar{m} = \max\{m : m \le n \wedge p \wedge 4n^2\phi(m)W_{\mathcal{U}}^2/\lambda^2\}$. Suppose that $\bar{m} > m_0 = n/\log(n\vee p)$, so that $\bar{m} = m_0\ell$ for some $\ell > 1$, since $\bar{m} \le n$ is finite. By definition, $\bar{m}$ satisfies $\bar{m} \le 4n^2\phi(\bar{m})W_{\mathcal{U}}^2/\lambda^2$. Insert the lower bound on λ, m_0 and $\bar{m} = m_0\ell$ in this inequality, and using Lemma 13 we obtain

$$\bar{m} = m_0\ell \le \frac{4n^2W_{\mathcal{U}}^2}{8W_{\mathcal{U}}^2 n\log(n\vee p)}\frac{\phi(m_0\ell)}{\phi(m_0)} \le \frac{n}{2\log(n\vee p)}\lceil\ell\rceil < \frac{n}{\log(n\vee p)}\ell = m_0\ell,$$

which is a contradiction. □

PROOF OF LEMMA 7 (Empirical sparsity). It is convenient to define:

1. the true rank scores, $a_i^*(u) = u - 1\{y_i \le x_i'\beta(u)\}$ for $i = 1,\dots,n$;
2. the estimated rank scores, $a_i(u) = u - 1\{y_i \le x_i'\widehat{\beta}(u)\}$ for $i = 1,\dots,n$;
3. the dual optimal rank scores, $\widehat{a}(u)$, that solve the dual program (C.2).

Let $\widehat{T}_u$ denote the support of $\widehat{\beta}(u)$, and $\widehat{s}_u = \|\widehat{\beta}(u)\|_0$. Let $\widetilde{x}_{i\widehat{T}_u} = (x_{ij}/\widehat{\sigma}_j,\ j \in \widehat{T}_u)'$, and $\widehat{\beta}_{\widehat{T}_u}(u) = (\widehat{\beta}_j(u),\ j \in \widehat{T}_u)'$. From the complementary slackness characterizations (C.3),

$$\sqrt{\widehat{s}_u} = \|\text{sign}(\widehat{\beta}_{\widehat{T}_u}(u))\| = \left\|\frac{n\mathbb{E}_n[\widetilde{x}_{i\widehat{T}_u}\widehat{a}_i(u)]}{\lambda\sqrt{u(1-u)}}\right\|. \tag{C.4}$$

Therefore, we can bound the number $\widehat{s}_u$ of nonzero components of $\widehat{\beta}(u)$ provided we can bound the empirical expectation in (C.4). This is achieved in the next step by combining the maximal inequalities and assumptions on the design matrix.

Using the triangle inequality in (C.4), write

$$\begin{aligned}\lambda\sqrt{\widehat{s}} \le \sup_{u\in\mathcal{U}}\{&\big(\|n\mathbb{E}_n[\widetilde{x}_{i\widehat{T}_u}(\widehat{a}_i(u) - a_i(u))]\|\\ &+ \|n\mathbb{E}_n[\widetilde{x}_{i\widehat{T}_u}(a_i(u) - a_i^*(u))]\|\\ &+ \|n\mathbb{E}_n[\widetilde{x}_{i\widehat{T}_u}a_i^*(u)]\|\big)\big(\sqrt{u(1-u)}\big)^{-1}\}.\end{aligned}$$

This leads to the inequality

$$\lambda\sqrt{\widehat{s}} \le \frac{W_{\mathcal{U}}}{\min_{j=1,\dots,p}\widehat{\sigma}_j}\Big(\sup_{u\in\mathcal{U}}\big\|n\mathbb{E}_n\big[x_{i\widehat{T}_u}\big(\widehat{a}_i(u)-a_i(u)\big)\big]\big\| + \sup_{u\in\mathcal{U}}\big\|n\mathbb{E}_n\big[x_{i\widehat{T}_u}\big(a_i(u)-a_i^*(u)\big)\big]\big\|\Big) + \sup_{u\in\mathcal{U}}\big\|n\mathbb{E}_n\big[\widetilde{x}_{i\widehat{T}_u}a_i^*(u)/\sqrt{u(1-u)}\big]\big\|.$$

Then we bound each of the three components in this display.

(a) To bound the first term, we observe that $\widehat{a}_i(u) \neq a_i(u)$ only if $y_i = x_i'\widehat{\beta}(u)$. By Lemma 9, the penalized quantile regression fit can interpolate at most $\widehat{s}_u \le \widehat{s}$ points with probability one uniformly over $u \in \mathcal{U}$. This implies that $\mathbb{E}_n[|\widehat{a}_i(u) - a_i(u)|^2] \le \widehat{s}/n$. Therefore,

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\big\|n\mathbb{E}_n\big[x_{i\widehat{T}_u}\big(\widehat{a}_i(u)-a_i(u)\big)\big]\big\| \\
&\quad\le n \sup_{\|\alpha\|_0\le\widehat{s},\|\alpha\|\le1}\sup_{u\in\mathcal{U}}\mathbb{E}_n[|\alpha'x_i||\widehat{a}_i(u)-a_i(u)|] \\
&\quad\le n \sup_{\|\alpha\|_0\le\widehat{s},\|\alpha\|\le1}\sqrt{\mathbb{E}_n[|\alpha'x_i|^2]}\sup_{u\in\mathcal{U}}\sqrt{\mathbb{E}_n[|\widehat{a}_i(u)-a_i(u)|^2]} \le \sqrt{n\phi(\widehat{s})\widehat{s}}.
\end{aligned}$$

(b) To bound the second term, note that

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\big\|n\mathbb{E}_n\big[x_{i\widehat{T}_u}\big(a_i(u)-a_i^*(u)\big)\big]\big\| \\
&\quad\le \sup_{u\in\mathcal{U}}\big\|\sqrt{n}\mathbb{G}_n\big(x_{i\widehat{T}_u}\big(a_i(u)-a_i^*(u)\big)\big)\big\| + \sup_{u\in\mathcal{U}}\big\|n\mathrm{E}\big[x_{i\widehat{T}_u}\big(a_i(u)-a_i^*(u)\big)\big]\big\| \\
&\quad\le \sqrt{n}\epsilon_1(r,\widehat{s}) + \sqrt{n}\epsilon_2(r,\widehat{s}),
\end{aligned}$$

where for $\psi_i(\beta,u) = (1\{y_i \le x_i'\beta\} - u)x_i$,

$$\begin{aligned}
\epsilon_1(r,m) &:= \sup_{u\in\mathcal{U},\beta\in R_u(r,m),\alpha\in\mathbb{S}(\beta)}|\mathbb{G}_n(\alpha'\psi_i(\beta,u)) - \mathbb{G}_n(\alpha'\psi_i(\beta(u),u))|, \\
\epsilon_2(r,m) &:= \sup_{u\in\mathcal{U},\beta\in R_u(r,m),\alpha\in\mathbb{S}(\beta)}\sqrt{n}|\mathrm{E}[\alpha'\psi_i(\beta,u)] - \mathrm{E}[\alpha'\psi_i(\beta(u),u)]|
\end{aligned} \tag{C.5}$$

and

$$\begin{aligned}
R_u(r,m) &:= \big\{\beta\in\mathbb{R}^p : \beta-\beta(u)\in A_u : \|\beta\|_0 \le m, \big\|J_u^{1/2}\big(\beta-\beta(u)\big)\big\| \le r\big\}, \\
\mathbb{S}(\beta) &:= \{\alpha\in\mathbb{R}^p : \|\alpha\| \le 1, \operatorname{support}(\alpha)\subseteq\operatorname{support}(\beta)\}.
\end{aligned} \tag{C.6}$$

By Lemma 12, there is a constant $A^1_{\varepsilon/2}$ such that

$$\sqrt{n}\epsilon_1(r,\widehat{s}) \le A^1_{\varepsilon/2}\sqrt{n\widehat{s}\log(n\vee p)}\sqrt{\phi(\widehat{s})}$$

with probability $1-\varepsilon/2$. By Lemma 10, we have $\sqrt{n}\epsilon_2(r,\widehat{s}) \le n(\mu(\widehat{s})/2)(r \wedge 1)$.

(c) To bound the last term, by Theorem 1 there exists a constant $A^0_{\varepsilon/2}$ such that with probability $1-\varepsilon/2$

$$\sup_{u\in\mathcal{U}} \|n\mathbb{E}_n[\widetilde{x}_i \widehat{T}_u a_i^*(u)/\sqrt{u(1-u)}]\| \le \sqrt{\widehat{s}}\Lambda \le \sqrt{\widehat{s}} A^0_{\varepsilon/2} W_{\mathcal{U}} \sqrt{n \log p},$$

where we used that $a_i^*(u) = u - 1\{u_i \le u\}$, $i = 1,\dots,n$, for $u_1,\dots,u_n$ i.i.d. uniform $(0,1)$.

Combining bounds in (a)–(c), using that $\min_{j=1,\dots,p} \widehat{\sigma}_j \ge 1/2$ by condition D.3 with probability $1-\gamma$, we have

$$\frac{\sqrt{\widehat{s}}}{W_{\mathcal{U}}} \le \mu(\widehat{s})\frac{n}{\lambda}(r \wedge 1) + \sqrt{\widehat{s}} K_\varepsilon \frac{\sqrt{n \log(n \vee p)\phi(\widehat{s})}}{\lambda},$$

with probability at least $1-\varepsilon-\gamma$, for $K_\varepsilon = 2(1 + A^0_{\varepsilon/2} + A^1_{\varepsilon/2})$. □

Next we control the linearization error ϵ_2 defined in (C.5).

LEMMA 10 (Controlling linearization error ϵ_2). *Under* D.1, D.2,

$$\epsilon_2(r,m) \le \sqrt{n}\sqrt{\varphi_{\max}(m)}\{1 \wedge (2[\bar{f}/\underline{f}^{1/2}]r)\} \qquad \textit{for all } r > 0 \textit{ and } m \le n.$$

PROOF. By definition

$$\epsilon_2(r,m) = \sup_{u\in\mathcal{U},\beta\in R_u(r,m),\alpha\in\mathbb{S}(\beta)} \sqrt{n}|\mathrm{E}[(\alpha' x_i)(1\{y_i \le x_i'\beta\} - 1\{y_i < x_i'\beta(u)\})]|.$$

By Cauchy–Schwarz, and using that $\varphi_{\max}(m) = \sup_{\|\alpha\|\le 1, \|\alpha\|_0 \le m} \mathrm{E}[|\alpha' x_i|^2]$,

$$\epsilon_2(r,m) \le \sqrt{n}\sqrt{\varphi_{\max}(m)} \sup_{u\in\mathcal{U},\beta\in R_u(r,m)} \sqrt{\mathrm{E}[\big(1\{y_i \le x_i'\beta\} - 1\{y_i < x_i'\beta(u)\}\big)^2]}.$$

Then, since for any $\beta \in R_u(r,m)$, $u \in \mathcal{U}$,

$$\begin{aligned}
&\mathrm{E}[\big(1\{y_i \le x_i'\beta\} - 1\{y_i < x_i'\beta(u)\}\big)^2] \\
&\quad \le \mathrm{E}[1\{|y_i - x_i'\beta(u)| \le |x_i'(\beta - \beta(u))|\}] \\
&\quad \le \mathrm{E}[(2\bar{f}|x_i'(\beta - \beta(u))|) \wedge 1] \le \{2\bar{f}(\mathrm{E}[|x_i'(\beta - \beta(u))|^2])^{1/2}\} \wedge 1
\end{aligned}$$

and $(\mathrm{E}[|x_i'(\beta - \beta(u))|^2])^{1/2} \le \|J_u^{1/2}(\beta - \beta(u))\|/\underline{f}^{1/2}$ by Lemma 4, the result follows. □

Next, we proceed to control the empirical error ϵ_1 defined in (C.5). We shall need the following preliminary result on the uniform L_2 covering numbers [33] of a relevant function class.

LEMMA 11. (1) *Consider a fixed subset* $T \subset \{1, 2, \ldots, p\}$, $|T| = m$. *The class of functions*

$$\mathcal{F}_T = \{\alpha'(\psi_i(\beta, u) - \psi_i(\beta(u), u)) : u \in \mathcal{U}, \alpha \in \mathbb{S}(\beta), \text{support}(\beta) \subseteq T\}$$

has a VC index bounded by cm for some universal constant c.

(2) *There are universal constants C and c such that for any* $m \leq n$ *the function class*

$$\mathcal{F}_m = \{\alpha'(\psi_i(\beta, u) - \psi_i(\beta(u), u)) : u \in \mathcal{U}, \beta \in \mathbb{R}^p, \|\beta\|_0 \leq m, \alpha \in \mathbb{S}(\beta)\}$$

has the the uniform covering numbers bounded as

$$\sup_Q N(\epsilon \|F_m\|_{Q,2}, \mathcal{F}_m, L_2(Q)) \leq C\left(\frac{16e}{\epsilon}\right)^{2(cm-1)}\left(\frac{ep}{m}\right)^m, \qquad \epsilon > 0.$$

PROOF. The proof involves standard combinatorial arguments and is relegated to [2]. □

LEMMA 12 (Controlling empirical error ϵ_1). *Under* D.1, D.2 *there exists a universal constant A such that with probability* $1 - \delta$

$$\epsilon_1(r, m) \leq A\delta^{-1/2}\sqrt{m \log(n \vee p)}\sqrt{\phi(m)} \qquad \textit{uniformly for all } r > 0 \textit{ and } m \leq n.$$

PROOF. By definition, $\epsilon_1(r, m) \leq \sup_{f \in \mathcal{F}_m} |\mathbb{G}_n(f)|$. From Lemma 11, the uniform covering number of $\mathcal{F}_m$ is bounded by $C(16e/\epsilon)^{2(cm-1)}(ep/m)^m$. Using Lemma 19 with $N = n$ and $\theta_m = p$, we have that uniformly in $m \leq n$, with probability at least $1 - \delta$

$$\begin{aligned}\sup_{f \in \mathcal{F}_m} |\mathbb{G}_n(f)| &\leq A\delta^{-1/2}\sqrt{m \log(n \vee p)} \\ &\quad \times \max\Big\{\sup_{f \in \mathcal{F}_m} \mathrm{E}[f^2]^{1/2}, \sup_{f \in \mathcal{F}_m} \mathbb{E}_n[f^2]^{1/2}\Big\}.\end{aligned} \tag{C.7}$$

By $|\alpha'(\psi_i(\beta, u) - \psi_i(\beta(u), u))| \leq |\alpha' x_i|$ and definition of $\phi(m)$

$$\mathbb{E}_n[f^2] \leq \mathbb{E}_n[|\alpha' x_i|^2] \leq \phi(m) \quad \text{and} \quad \mathrm{E}[f^2] \leq \mathrm{E}[|\alpha' x_i|^2] \leq \phi(m). \tag{C.8}$$

Combining (C.8) with (C.7), we obtain the result. □

(c) The next lemma provides a bound on maximum k-sparse eigenvalues, which we used in some of the derivations presented earlier.

LEMMA 13. *Let M be a semi-definite positive matrix and* $\phi_M(k) = \sup\{\alpha' \times M\alpha : \alpha \in \mathbb{R}^p, \|\alpha\| = 1, \|\alpha\|_0 \leq k\}$. *For any integers k and* ℓk *with* $\ell \geq 1$, *we have* $\phi_M(\ell k) \leq \lceil \ell \rceil \phi_M(k)$.

PROOF. See [2]. □

APPENDIX D: PROOF OF THEOREM 4

PROOF OF THEOREM 4. See [2]. □

APPENDIX E: PROOF OF LEMMA 8 (USED IN THEOREM 5)

PROOF OF LEMMA 8 (Sparse identifiability and control of empirical error). The proof of claim (3.17) of this lemma identically follows the proof of claim (3.7) of Lemma 4, given in Appendix B, after replacing A_u with $\widetilde{A}_u$. Next, we bound the empirical error

$$\begin{aligned}
&\sup_{u\in\mathcal{U},\delta\in\widetilde{A}_u(\widetilde{m}),\delta\neq 0}\frac{|\epsilon_u(\delta)|}{\|\delta\|}\\
&\quad\le \sup_{u\in\mathcal{U},\delta\in\widetilde{A}_u(\widetilde{m}),\delta\neq 0}\frac{1}{\|\delta\|\sqrt{n}}\left|\int_0^1 \delta'\mathbb{G}_n\big(\psi_i\big(\beta(u)+\gamma\delta,u\big)\big)\,d\gamma\right| \qquad \text{(E.1)}\\
&\quad\le \frac{1}{\sqrt{n}}\epsilon_3(\widetilde{m}),
\end{aligned}$$

where $\epsilon_3(\widetilde{m}) := \sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}}|\mathbb{G}_n(f)|$ and the class of functions $\widetilde{\mathcal{F}}_{\widetilde{m}}$ is defined in Lemma 14. The result follows from the bound on $\epsilon_3(\widetilde{m})$ holding uniformly in $\widetilde{m}\le n$ given in Lemma 15. □

Next, we control the empirical error ϵ_3 defined in (E.1) for $\widetilde{\mathcal{F}}_{\widetilde{m}}$ defined below. We first bound uniform covering numbers of $\widetilde{\mathcal{F}}_{\widetilde{m}}$.

LEMMA 14. *Consider a fixed subset $T\subset\{1,2,\dots,p\}$, $T_u=\operatorname{support}(\beta(u))$ such that $|T\setminus T_u|\le\widetilde{m}$ and $|T_u|\le s$ for some $u\in\mathcal{U}$. The class of functions*

$$\mathcal{F}_{T,u}=\big\{\alpha' x_i(1\{y_i\le x_i'\beta\}-u):\alpha\in\mathbb{S}(\beta),\operatorname{support}(\beta)\subseteq T\big\}$$

has a VC index bounded by $c(\widetilde{m}+s)+2$. The class of functions

$$\widetilde{\mathcal{F}}_{\widetilde{m}}=\big\{\mathcal{F}_{T,u}:u\in\mathcal{U},T\subset\{1,2,\dots,p\},|T\setminus T_u|\le\widetilde{m}\big\},$$

obeys, for some universal constants C and c and each $\epsilon>0$,

$$\sup_Q N(\epsilon\|\widetilde{F}_{\widetilde{m}}\|_{Q,2},\widetilde{\mathcal{F}}_{\widetilde{m}},L_2(Q))\le C(32e/\epsilon)^{4(c(\widetilde{m}+s)+2)}p^{2\widetilde{m}}\left|\bigcup_{u\in\mathcal{U}}T_u\right|^{2s}.$$

PROOF. The proof involves standard combinatorial arguments and is relegated to [2]. □

LEMMA 15 (Controlling empirical error ϵ_3). *Suppose that D.1 holds and* $|\bigcup_{u\in\mathcal{U}} T_u| \le n$. *There exists a universal constant A such that with probability at least* $1-\delta$,

$$\epsilon_3(\widetilde{m}) := \sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}} |\mathbb{G}_n(f)| \le A\delta^{-1/2}\sqrt{(\widetilde{m}\log(n\vee p)+s\log n)\phi(\widetilde{m}+s)}$$

for all $\widetilde{m}\le n$.

PROOF. Lemma 14 bounds the uniform covering number of $\widetilde{\mathcal{F}}_{\widetilde{m}}$. Using Lemma 19 with $N\le 2n$, $m=\widetilde{m}+s$ and $\theta_m = p^{2([m-s]/m)}\cdot n^{2(s/m)} = p^{2(\widetilde{m}/[\widetilde{m}+s])}\cdot n^{2(s/[\widetilde{m}+s])}$, we conclude that uniformly in $0\le\widetilde{m}\le n$

$$\begin{aligned}\sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}} |\mathbb{G}_n(f)| &\le A\delta^{-1/2}\sqrt{(\widetilde{m}+s)\log(n\vee\theta_m)}\\ &\quad\times\max\Big\{\sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}}\mathrm{E}[f^2]^{1/2},\ \sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}}\mathbb{E}_n[f^2]^{1/2}\Big\}\\ &\le A'\delta^{-1/2}\sqrt{\widetilde{m}\log(n\vee p)+s\log n}\\ &\quad\times\max\Big\{\sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}}\mathrm{E}[f^2]^{1/2},\ \sup_{f\in\widetilde{\mathcal{F}}_{\widetilde{m}}}\mathbb{E}_n[f^2]^{1/2}\Big\}\end{aligned} \tag{E.2}$$

with probability at least $1-\delta$. The result follows, since for any $f\in\widetilde{\mathcal{F}}_{\widetilde{m}}$, the corresponding vector α obeys $\|\alpha\|_0\le\widetilde{m}+s$, so that $\mathbb{E}_n[f^2]\le\mathbb{E}_n[|\alpha' x_i|^2]\le\phi(\widetilde{m}+s)$ and $\mathrm{E}[f^2]\le\mathrm{E}[|\alpha' x_i|^2]\le\phi(\widetilde{m}+s)$ by definition of $\phi(\widetilde{m}+s)$. □

APPENDIX F: MAXIMAL INEQUALITIES FOR A COLLECTION OF EMPIRICAL PROCESSES

The main results here are Lemma 16 and Lemma 19, used in the proofs of Theorem 1 and Theorems 3 and 5, respectively. Lemma 19 gives a maximal inequality that controls the empirical process uniformly over a collection of classes of functions using class-dependent bounds. We need this lemma because the standard maximal inequalities applied to the union of function classes yield a single class-independent bound that is too large for our purposes. We prove Lemma 19 by first stating Lemma 16, giving a bound on tail probabilities of a separable sub-Gaussian process, stated in terms of uniform covering numbers. Here we want to explicitly trace the impact of covering numbers on the tail probability, since these covering numbers grow rapidly under increasing parameter dimension and thus help to tighten the probability bound. Using the symmetrization approach, we then obtain Lemma 18, giving a bound on tail probabilities of a general separable empirical process, also stated in terms of uniform covering numbers. Finally, given a growth rate on the covering numbers, we obtain Lemma 19.

LEMMA 16 (Exponential inequality for sub-Gaussian process). *Consider any linear zero-mean separable process* $\{\mathbb{G}(f): f \in \mathcal{F}\}$, *whose index set* $\mathcal{F}$ *includes zero, is equipped with a* $L_2(P)$ *norm, and has envelope* F. *Suppose further that the process is sub-Gaussian, namely for each* $g \in \mathcal{F} - \mathcal{F}: \mathbb{P}\{|\mathbb{G}(g)| > \eta\} \leq 2\exp(-\frac{1}{2}\eta^2/D^2\|g\|_{P,2}^2)$ *for any* $\eta > 0$, *with* D *a positive constant; and suppose that we have the following upper bound on the* $L_2(P)$ *covering numbers for* $\mathcal{F}$:

$$N(\epsilon\|F\|_{P,2}, \mathcal{F}, L_2(P)) \leq n(\epsilon, \mathcal{F}, P) \qquad \text{for each } \epsilon > 0,$$

where $n(\epsilon, \mathcal{F}, P)$ *is increasing in* $1/\epsilon$, *and* $\epsilon\sqrt{\log n(\epsilon, \mathcal{F}, P)} \to 0$ *as* $1/\epsilon \to \infty$ *and is decreasing in* $1/\epsilon$. *Then for* $K > D$, *for some universal constant* $c < 30$, $\rho(\mathcal{F}, P) := \sup_{f\in\mathcal{F}} \|f\|_{P,2}/\|F\|_{P,2}$,

$$\mathbb{P}\left\{\frac{\sup_{f\in\mathcal{F}}|\mathbb{G}(f)|}{\|F\|_{P,2}\int_0^{\rho(\mathcal{F},P)/4}\sqrt{\log n(x, \mathcal{F}, P)}\,dx} > cK\right\} \leq \int_0^{\rho(\mathcal{F},P)/2} \epsilon^{-1} n(\epsilon, \mathcal{F}, P)^{-\{(K/D)^2-1\}}\,d\epsilon.$$

The result of Lemma 16 is in spirit of the Talagrand tail inequality for Gaussian processes. Our result is less sharp than Talagrand's result in the Gaussian case (by a log factor), but it applies to more general sub-Gaussian processes.

In order to prove a bound on tail probabilities of a general separable empirical process, we need to go through a symmetrization argument. Since we use a data-dependent threshold, we need an appropriate extension of the classical symmetrization lemma to allow for this. Let us call a threshold function $x: \mathbb{R}^n \mapsto \mathbb{R}$ k-sub-exchangeable if, for any $v, w \in \mathbb{R}^n$ and any vectors $\widetilde{v}, \widetilde{w}$ created by the pairwise exchange of the components in v with components in w, we have that $x(\widetilde{v}) \vee x(\widetilde{w}) \geq [x(v) \vee x(w)]/k$. Several functions satisfy this property, in particular $x(v) = \|v\|$ with $k = \sqrt{2}$ and constant functions with $k = 1$. The following result generalizes the standard symmetrization lemma for probabilities (Lemma 2.3.7 of [33]) to the case of a random threshold x that is sub-exchangeable.

LEMMA 17 (Symmetrization with data-dependent thresholds). *Consider arbitrary independent stochastic processes* $Z_1, \ldots, Z_n$ *and arbitrary functions* $\mu_1, \ldots, \mu_n: \mathcal{F} \mapsto \mathbb{R}$. *Let* $x(Z) = x(Z_1, \ldots, Z_n)$ *be a* k*-sub-exchangeable random variable and for any* $\tau \in (0, 1)$ *let* q_τ *denote the* τ *quantile of* $x(Z)$, $\bar{p}_\tau := P(x(Z) \leq q_\tau) \geq \tau$ *and* $p_\tau := P(x(Z) < q_\tau) \leq \tau$. *Then*

$$P\left(\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}} > x_0 \vee x(Z)\right) \leq \frac{4}{\bar{p}_\tau} P\left(\left\|\sum_{i=1}^n \varepsilon_i(Z_i - \mu_i)\right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z)}{4k}\right) + p_\tau,$$

where x_0 *is a constant such that* $\inf_{f\in\mathcal{F}} P(|\sum_{i=1}^n Z_i(f)| \leq \frac{x_0}{2}) \geq 1 - \frac{\bar{p}_\tau}{2}$.

Note that we can recover the classical symmetrization lemma for fixed thresholds by setting $k = 1$, $\bar{p}_\tau = 1$ and $p_\tau = 0$.

LEMMA 18 (Exponential inequality for separable empirical process). *Consider a separable empirical process* $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n\{f(Z_i) - \mathrm{E}[f(Z_i)]\}$ *and the empirical measure* $\mathbb{P}_n$ *for* $Z_1, \ldots, Z_n$, *an underlying i.i.d. data sequence. Let* $K > 1$ *and* $\tau \in (0, 1)$ *be constants, and* $e_n(\mathcal{F}, \mathbb{P}_n) = e_n(\mathcal{F}, Z_1, \ldots, Z_n)$ *be a k-sub-exchangeable random variable, such that*

$$\|F\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F},\mathbb{P}_n)/4}\sqrt{\log n(\epsilon, \mathcal{F}, \mathbb{P}_n)}\,d\epsilon \le e_n(\mathcal{F}, \mathbb{P}_n)$$

and

$$\sup_{f\in\mathcal{F}}\operatorname*{var}_{\mathbb{P}} f \le \frac{\tau}{2}(4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2$$

for the same constant $c > 0$ *as in Lemma 16, then*

$$\mathbb{P}\Big\{\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)| \ge 4kcKe_n(\mathcal{F}, \mathbb{P}_n)\Big\}$$
$$\le \frac{4}{\tau}\mathrm{E}_{\mathbb{P}}\left(\left[\int_0^{\rho(\mathcal{F},\mathbb{P}_n)/2}\epsilon^{-1}n(\epsilon, \mathcal{F}, \mathbb{P}_n)^{-\{K^2-1\}}\,d\epsilon\right]\wedge 1\right) + \tau.$$

Finally, our main result in this section is as follows.

LEMMA 19 (Maximal inequality for a collection of empirical processes). *Consider a collection of separable empirical processes*

$$\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n\{f(Z_i) - \mathrm{E}[f(Z_i)]\},$$

where $Z_1, \ldots, Z_n$ *is an underlying i.i.d. data sequence, defined over function classes* $\mathcal{F}_m, m = 1, \ldots, N$ *with envelopes* $F_m = \sup_{f\in\mathcal{F}_m}|f(x)|, m = 1, \ldots, N$, *and with upper bounds on the uniform covering numbers of* $\mathcal{F}_m$ *given for all m by*

$$n(\epsilon, \mathcal{F}_m, \mathbb{P}_n) = (N \vee n \vee \theta_m)^m(\omega/\epsilon)^{\upsilon m}, \qquad 0 < \epsilon < 1,$$

with some constants $\omega > 1$, $\upsilon > 1$ *and* $\theta_m \ge \theta_0$. *For a constant* $C := (1 + \sqrt{2\upsilon})/4$ *set*

$$e_n(\mathcal{F}_m, \mathbb{P}_n) = C\sqrt{m\log(N \vee n \vee \theta_m \vee \omega)}\max\Big\{\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P},2}, \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\Big\}.$$

Then, for any $\delta \in (0, 1/6)$, *and any constant* $K \ge \sqrt{2/\delta}$ *we have*

$$\sup_{f\in\mathcal{F}_m}|\mathbb{G}_n(f)| \le 4\sqrt{2}cKe_n(\mathcal{F}_m, \mathbb{P}_n) \qquad \text{for all } m \le N,$$

with probability at least $1 - \delta$, *provided that* $N \vee n \vee \theta_0 \ge 3$; *the constant c is the same as in Lemma 16.*

PROOF OF LEMMA 16. The strategy of the proof is similar to the proof of Lemma 19.34 in [31], page 286, given for the expectation of a supremum of a process; here we instead bound tail probabilities and also compute all constants explicitly.

Step 1. There exists a sequence of nested partitions of $\mathcal{F}$, $\{(\mathcal{F}_{qi}, i = 1, \ldots, N_q), q = q_0, q_0 + 1, \ldots\}$ where the qth partition consists of sets of $L_2(P)$ radius at most $\|F\|_{P,2}2^{-q}$, and q_0 is the largest positive integer such that $2^{-q_0} \leq \rho(\mathcal{F}, P)/4$ so that $q_0 \geq 2$. The existence of such a partition follows from a standard argument, for example, [31], page 286.

Let f_{qi} be an arbitrary point of $\mathcal{F}_{qi}$. Set $\pi_q(f) = f_{qi}$ if $f \in \mathcal{F}_{qi}$. By separability of the process, we can replace $\mathcal{F}$ by $\bigcup_{q,i} f_{qi}$, since the supremum norm of the process can be computed by taking this set only. In this case, we can decompose $f - \pi_{q_0}(f) = \sum_{q=q_0+1}^{\infty}(\pi_q(f) - \pi_{q-1}(f))$. Hence by linearity $\mathbb{G}(f) - \mathbb{G}(\pi_{q_0}(f)) = \sum_{q=q_0+1}^{\infty} \mathbb{G}(\pi_q(f) - \pi_{q-1}(f))$, so that

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} |\mathbb{G}(f)| > \sum_{q=q_0}^{\infty} \eta_q\right\} \leq \sum_{q=q_0+1}^{\infty} \mathbb{P}\Big\{\max_f |\mathbb{G}\big(\pi_q(f) - \pi_{q-1}(f)\big)| > \eta_q\Big\} + \mathbb{P}\Big\{\max_f |\mathbb{G}(\pi_{q_0}(f))| > \eta_{q_0}\Big\},$$

for constants η_q chosen below.

Step 2. By construction of the partition sets $\|\pi_q(f) - \pi_{q-1}(f)\|_{P,2} \leq 2 \times \|F\|_{P,2}2^{-(q-1)} \leq 4\|F\|_{P,2}2^{-q}$, for $q \geq q_0 + 1$. Setting $\eta_q = 8K\|F\|_{P,2}2^{-q} \times \sqrt{\log N_q}$, using sub-Gaussianity, setting $K > D$, using that $2\log N_q \geq \log N_q \times N_{q-1} \geq \log n_q$, using that $q \mapsto \log n_q$ is increasing in q, and $2^{-q_0} \leq \rho(\mathcal{F}, P)/4$, we obtain

$$\begin{aligned}
&\sum_{q=q_0+1}^{\infty} \mathbb{P}\Big\{\max_f |\mathbb{G}\big(\pi_q(f) - \pi_{q-1}(f)\big)| > \eta_q\Big\} \\
&\quad\leq \sum_{q=q_0+1}^{\infty} N_q N_{q-1} 2\exp\big(-\eta_q^2/(4D\|F\|_{P,2}2^{-q})^2\big) \\
&\quad\leq \sum_{q=q_0+1}^{\infty} N_q N_{q-1} 2\exp\big(-(K/D)^2 2\log N_q\big) \\
&\quad\leq \sum_{q=q_0+1}^{\infty} 2\exp\big(-\{(K/D)^2 - 1\}\log n_q\big) \\
&\quad\leq \int_{q_0}^{\infty} 2\exp\big(-\{(K/D)^2 - 1\}\log n_q\big)\, dq \\
&\quad= \int_0^{\rho(\mathcal{F},P)/4} (x\ln 2)^{-1} 2n(x, \mathcal{F}, P)^{-\{(K/D)^2-1\}}\, dx.
\end{aligned}$$

By Jensen's inequality, we have $\sqrt{\log N_q} \le a_q := \sum_{j=q_0}^{q} \sqrt{\log n_j}$, so that we obtain $\sum_{q=q_0+1}^{\infty} \eta_q \le 8 \sum_{q=q_0+1}^{\infty} K \|F\|_{P,2} 2^{-q} a_q$. Letting $b_q = 2 \cdot 2^{-q}$, noting $a_{q+1} - a_q = \sqrt{\log n_{q+1}}$ and $b_{q+1} - b_q = -2^{-q}$, we get using summation by parts

$$\begin{aligned}
\sum_{q=q_0+1}^{\infty} 2^{-q} a_q &= -\sum_{q=q_0+1}^{\infty} (b_{q+1} - b_q) a_q \\
&= -a_q b_q|_{q_0+1}^{\infty} + \sum_{q=q_0+1}^{\infty} b_{q+1}(a_{q+1} - a_q) \\
&= 2 \cdot 2^{-(q_0+1)} \sqrt{\log n_{q_0+1}} + \sum_{q=q_0+1}^{\infty} 2 \cdot 2^{-(q+1)} \sqrt{\log n_{q+1}} \\
&= 2 \sum_{q=q_0+1}^{\infty} 2^{-q} \sqrt{\log n_q},
\end{aligned}$$

where we use the assumption that $2^{-q}\sqrt{\log n_q} \to 0$ as $q \to \infty$, so that

$$-a_q b_q|_{q_0+1}^{\infty} = 2 \cdot 2^{-(q_0+1)} \sqrt{\log n_{q_0+1}}.$$

Using that $2^{-q}\sqrt{\log n_q}$ is decreasing in q by assumption,

$$2 \sum_{q=q_0+1}^{\infty} 2^{-q} \sqrt{\log n_q} \le 2 \int_{q_0}^{\infty} 2^{-q} \sqrt{\log n(2^{-q}, \mathcal{F}, P)}\, dq.$$

Using a change of variables and that $2^{-q_0} \le \rho(\mathcal{F}, P)/4$, we finally conclude that

$$\sum_{q=q_0+1}^{\infty} \eta_q \le K \|F\|_{P,2} \frac{16}{\log 2} \int_0^{\rho(\mathcal{F},P)/4} \sqrt{\log n(x, \mathcal{F}, P)}\, dx.$$

Step 3. Letting $\eta_{q_0} = K \|F\|_{P,2} \rho(\mathcal{F}, P) \sqrt{2 \log N_{q_0}}$, recalling that $N_{q_0} = n_{q_0}$, using that $\|\pi_{q_0}(f)\|_{P,2} \le \|F\|_{P,2}$ and sub-Gaussianity, we conclude

$$\begin{aligned}
&\mathbb{P}\Big\{\max_f |\mathbb{G}(\pi_{q_0}(f))| > \eta_{q_0}\Big\} \\
&\quad \le n_q 2 \exp\big(-(K/D)^2 \log n_q\big) \le 2 \exp\big(-\{(K/D)^2 - 1\} \log n_q\big) \\
&\quad \le \int_{q_0-1}^{q_0} 2 \exp\big(-\{(K/D)^2 - 1\} \log n_q\big)\, dq \\
&\quad = \int_{\rho(\mathcal{F},P)/4}^{\rho(\mathcal{F},P)/2} (x \ln 2)^{-1} 2 n(x, \mathcal{F}, P)^{-\{(K/D)^2 - 1\}}\, dx.
\end{aligned}$$

Also, since $n_{q_0} = n(2^{-q_0}, \mathcal{F}, P)$, $2^{-q_0} \leq \rho(\mathcal{F}, P)/4$ and $n(x, \mathcal{F}, P)$ is increasing in $1/x$, we obtain $\eta_{q_0} \leq 4\sqrt{2}K\|F\|_{P,2}\int_0^{\rho(\mathcal{F},P)/4}\sqrt{\log n(x, \mathcal{F}, P)}\,dx$.

Step 4. Finally, adding the bounds on tail probabilities from Steps 2 and 3 we obtain the tail bound stated in the main text. Further, adding bounds on η_q from Steps 2 and 3, and using $c = 16/\log 2 + 4\sqrt{2} < 30$, we obtain $\sum_{q=q_0}^{\infty}\eta_q \leq cK\|F\|_{P,2}\int_0^{\rho(\mathcal{F},P)/4}\sqrt{\log n(x, \mathcal{F}, P)}\,dx$. □

PROOF OF LEMMA 17. The proof proceeds analogously to the proof of Lemma 2.3.7 (page 112) in [33] with the necessary adjustments. Letting q_τ be the τ quantile of $x(Z)$ we have

$$P\left\{\left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}} > x_0 \vee x(Z)\right\} \leq P\left\{x(Z) \geq q_\tau, \left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}} > x_0 \vee x(Z)\right\} + P\{x(Z) < q_\tau\}.$$

Next we bound the first term of the expression above. Let $Y = (Y_1, \ldots, Y_n)$ be an independent copy of $Z = (Z_1, \ldots, Z_n)$, suitably defined on a product space. Fix a realization of Z such that $x(Z) \geq q_\tau$ and $\|\sum_{i=1}^n Z_i\|_{\mathcal{F}} > x_0 \vee x(Z)$. Therefore, $\exists f_Z \in \mathcal{F}$ such that $|\sum_{i=1}^n Z_i(f_Z)| > x_0 \vee x(Z)$. Conditional on such a Z and using the triangular inequality, we have that

$$\begin{aligned} &P_Y\left\{x(Y) \leq q_\tau, \left|\sum_{i=1}^n Y_i(f_Z)\right| \leq \frac{x_0}{2}\right\} \\ &\quad\leq P_Y\left\{\left|\sum_{i=1}^n (Y_i - Z_i)(f_Z)\right| > \frac{x_0 \vee x(Z) \vee x(Y)}{2}\right\} \\ &\quad\leq P_Y\left\{\left\|\sum_{i=1}^n (Y_i - Z_i)\right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2}\right\}. \end{aligned}$$

By definition of x_0, we have $\inf_{f\in\mathcal{F}} P\{|\sum_{i=1}^n Y_i(f)| \leq \frac{x_0}{2}\} \geq 1 - \bar{p}_\tau/2$. Since $P_Y\{x(Y) \leq q_\tau\} = \bar{p}_\tau$, by Bonferroni inequality we have that the left-hand side is bounded from below by $\bar{p}_\tau - \bar{p}_\tau/2 = \bar{p}_\tau/2$. Therefore, over the set $\{Z : x(Z) \geq q_\tau, \|\sum_{i=1}^n Z_i\|_{\mathcal{F}} > x_0 \vee x(Z)\}$ we have

$$\frac{\bar{p}_\tau}{2} \leq P_Y\left\{\left\|\sum_{i=1}^n (Y_i - Z_i)\right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2}\right\}.$$

Integrating over Z, we obtain

$$\begin{aligned} &\frac{\bar{p}_\tau}{2} P\left\{x(Z) \geq q_\tau, \left\|\sum_{i=1}^n Z_i\right\|_{\mathcal{F}} > x_0 \vee x(Z)\right\} \\ &\quad\leq P_Z P_Y\left\{\left\|\sum_{i=1}^n (Y_i - Z_i)\right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2}\right\}. \end{aligned}$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ be an independent sequence of Rademacher random variables. Given $\varepsilon_1, \ldots, \varepsilon_n$, set $(\widetilde{Y}_i = Y_i, \widetilde{Z}_i = Z_i)$ if $\varepsilon_i = 1$ and $(\widetilde{Y}_i = Z_i, \widetilde{Z}_i = Y_i)$ if $\varepsilon_i = -1$. That is, we create vectors $\widetilde{Y}$ and $\widetilde{Z}$ by pairwise exchanging their components; by construction, conditional on each $\varepsilon_1, \ldots, \varepsilon_n$, $(\widetilde{Y}, \widetilde{Z})$ has the same distribution as (Y, Z). Therefore,

$$P_Z P_Y \left\{ \left\| \sum_{i=1}^n (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2} \right\}$$
$$= \mathrm{E}_\varepsilon P_Z P_Y \left\{ \left\| \sum_{i=1}^n (\widetilde{Y}_i - \widetilde{Z}_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(\widetilde{Z}) \vee x(\widetilde{Y})}{2} \right\}.$$

By $x(\cdot)$ being k-sub-exchangeable, and since $\varepsilon_i(Y_i - Z_i) = (\widetilde{Y}_i - \widetilde{Z}_i)$, we have that

$$\mathrm{E}_\varepsilon P_Z P_Y \left\{ \left\| \sum_{i=1}^n (\widetilde{Y}_i - \widetilde{Z}_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(\widetilde{Z}) \vee x(\widetilde{Y})}{2} \right\}$$
$$\leq \mathrm{E}_\varepsilon P_Z P_Y \left\{ \left\| \sum_{i=1}^n \varepsilon_i (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z) \vee x(Y)}{2k} \right\}.$$

By the triangular inequality and removing $x(Y)$ or $x(Z)$, the latter is bounded by

$$P \left\{ \left\| \sum_{i=1}^n \varepsilon_i (Y_i - \mu_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Y)}{4k} \right\}$$
$$+ P \left\{ \left\| \sum_{i=1}^n \varepsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} > \frac{x_0 \vee x(Z)}{4k} \right\}. \qquad \square$$

PROOF OF LEMMA 18. Let $\mathbb{G}_n^o(f) = n^{-1/2} \sum_{i=1}^n \{\varepsilon_i f(Z_i)\}$ be the symmetrized empirical process, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables, that is, $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$, which are independent of $Z_1, \ldots, Z_n$. By the Chebyshev's inequality and the assumption on $e_n(\mathcal{F}, \mathbb{P}_n)$, we have for the constant τ fixed in the statement of the lemma

$$P\big(|\mathbb{G}_n(f)| > 4kcKe_n(\mathcal{F}, \mathbb{P}_n)\big) \leq \frac{\sup_f \mathrm{var}_{\mathbb{P}} \mathbb{G}_n(f)}{(4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2} = \frac{\sup_{f \in \mathcal{F}} \mathrm{var}_{\mathbb{P}} f}{(4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2}$$
$$\leq \tau/2.$$

Therefore, by the symmetrization Lemma 17, we obtain

$$\mathbb{P}\Big\{ \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| > 4kcKe_n(\mathcal{F}, \mathbb{P}_n) \Big\} \leq \frac{4}{\tau} \mathbb{P}\Big\{ \sup_{f \in \mathcal{F}} |\mathbb{G}_n^o(f)| > cKe_n(\mathcal{F}, \mathbb{P}_n) \Big\} + \tau.$$

We then condition on the values of $Z_1, \ldots, Z_n$, denoting the conditional probability measure as $\mathbb{P}_\varepsilon$. Conditional on $Z_1, \ldots, Z_n$, by the Hoeffding inequality

the symmetrized process $\mathbb{G}_n^o$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$ norm, namely, for $g \in \mathcal{F} - \mathcal{F}$, $\mathbb{P}_\varepsilon\{\mathbb{G}_n^o(g) > x\} \leq 2\exp(-x^2/[2\|g\|_{\mathbb{P}_n,2}^2])$. Hence, by Lemma 16 with $D = 1$, we can bound

$$\mathbb{P}_\varepsilon\Big\{\sup_{f\in\mathcal{F}}|\mathbb{G}_n^o(f)| \geq cKe_n(\mathcal{F},\mathbb{P}_n)\Big\} \leq \left[\int_0^{\rho(\mathcal{F},\mathbb{P}_n)/2} \epsilon^{-1} n(\epsilon,\mathcal{F},P)^{-\{K^2-1\}}\,d\epsilon\right] \wedge 1.$$

The result follows from taking the expectation over $Z_1,\ldots,Z_n$. □

PROOF OF LEMMA 19. Step 1 (Main step). In this step, we prove the main result. First, we observe that the bound $\epsilon \mapsto n(\epsilon,\mathcal{F}_m,\mathbb{P}_n)$ satisfies the monotonicity hypotheses of Lemma 18 uniformly in $m \leq N$.

Second, recall $e_n(\mathcal{F}_m,\mathbb{P}_n) := C\sqrt{m\log(N\vee n\vee\theta_m\vee\omega)}\max\{\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P},2}, \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\}$ for $C = (1+\sqrt{2\upsilon})/4$. Note that $\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}$ is $\sqrt{2}$-sub-exchangeable and $\rho(\mathcal{F}_m,\mathbb{P}_n) := \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}/\|F_m\|_{\mathbb{P}_n,2} \geq 1/\sqrt{n}$ by Step 2 below. Thus, uniformly in $m \leq N$,

$$\begin{aligned}
&\|F_m\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F}_m,\mathbb{P}_n)/4}\sqrt{\log n(\epsilon,\mathcal{F}_m,\mathbb{P}_n)}\,d\epsilon\\
&\quad\leq \|F_m\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F}_m,\mathbb{P}_n)/4}\sqrt{m\log(N\vee n\vee\theta_m)+\upsilon m\log(\omega/\epsilon)}\,d\epsilon\\
&\quad\leq (1/4)\sqrt{m\log(N\vee n\vee\theta_m)}\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\\
&\qquad+\|F_m\|_{\mathbb{P}_n,2}\int_0^{\rho(\mathcal{F}_m,\mathbb{P}_n)/4}\sqrt{\upsilon m\log(\omega/\epsilon)}\,d\epsilon\\
&\quad\leq \sqrt{m\log(N\vee n\vee\theta_m\vee\omega)}\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}(1+\sqrt{2\upsilon})/4\\
&\quad\leq e_n(\mathcal{F}_m,\mathbb{P}_n),
\end{aligned}$$

which follows by

$$\begin{aligned}
\int_0^\rho\sqrt{\log(\omega/\epsilon)}\,d\epsilon &\leq \left(\int_0^\rho 1\,d\epsilon\right)^{1/2}\left(\int_0^\rho\log(\omega/\epsilon)\,d\epsilon\right)^{1/2}\\
&\leq \rho\sqrt{2\log(n\vee\omega)}\qquad\text{for } 1/\sqrt{n}\leq\rho\leq 1.
\end{aligned}$$

Third, for any $K \geq \sqrt{2/\delta} > 1$ we have $(K^2-1) \geq 1/\delta$, and let $\tau_m = \delta/(4m\log(N\vee n\vee\theta_0))$. Recall that $4\sqrt{2}cC > 4$ where $4 < c < 30$ is defined in Lemma 16. Note that for any $m \leq N$ and $f \in \mathcal{F}_m$, we have by Chebyshev's in-

equality

$$P\big(|\mathbb{G}_n(f)| > 4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n)\big) \le \frac{\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P},2}}{(4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n))^2} \le \frac{\delta/2}{(4\sqrt{2}cC)^2 m\log(N\vee n\vee\theta_0)} \le \tau_m/2.$$

By Lemma 18 with our choice of τ_m, $m\le N$, $\omega>1$, $\upsilon>1$ and $\rho(\mathcal{F}_m,\mathbb{P}_n)\le 1$,

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{f\in\mathcal{F}_m}|\mathbb{G}_n(f)| > 4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n),\exists m\le N\Big\}\\
&\le \sum_{m=1}^N \mathbb{P}\Big\{\sup_{f\in\mathcal{F}_m}|\mathbb{G}_n(f)| > 4\sqrt{2}cKe_n(\mathcal{F}_m,\mathbb{P}_n)\Big\}\\
&\le \sum_{m=1}^N\Big[\frac{4(N\vee n\vee\theta_m)^{-m/\delta}}{\tau_m}\int_0^{1/2}(\omega/\epsilon)^{(-\upsilon m/\delta)+1}\,d\epsilon+\tau_m\Big]\\
&\le 4\sum_{m=1}^N\frac{(N\vee n\vee\theta_m)^{-m/\delta}}{\tau_m}\frac{1}{\upsilon m/\delta}+\sum_{m=1}^N\tau_m\\
&< 16\frac{(N\vee n\vee\theta_0)^{-1/\delta}}{1-(N\vee n\vee\theta_0)^{-1/\delta}}\log(N\vee n\vee\theta_0)+\frac{\delta}{4}\frac{(1+\log N)}{\log(N\vee n\vee\theta_0)}\le\delta,
\end{aligned}$$

where the last inequality follows by $N\vee n\vee\theta_0\ge 3$ and $\delta\in(0,1/6)$.

Step 2 (Auxiliary calculations). To establish that $\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}$ is $\sqrt{2}$-sub-exchangeable, let $\widetilde{Z}$ and $\widetilde{Y}$ be created by exchanging any components in Z with corresponding components in Y. Then

$$\begin{aligned}
&\sqrt{2}\Big(\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(\widetilde{Z}),2}\vee\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(\widetilde{Y}),2}\Big)\\
&\ge\Big(\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(\widetilde{Z}),2}+\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(\widetilde{Y}),2}\Big)^{1/2}\\
&\ge\Big(\sup_{f\in\mathcal{F}_m}\mathbb{E}_n[f(\widetilde{Z}_i)^2]+\mathbb{E}_n[f(\widetilde{Y}_i)^2]\Big)^{1/2}\\
&=\Big(\sup_{f\in\mathcal{F}_m}\mathbb{E}_n[f(Z_i)^2]+\mathbb{E}_n[f(Y_i)^2]\Big)^{1/2}\\
&\ge\Big(\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(Z),2}\vee\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(Y),2}\Big)^{1/2}\\
&=\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(Z),2}\vee\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(Y),2}.
\end{aligned}$$

Next we show that $\rho(\mathcal{F}_m, \mathbb{P}_n) := \sup_{f\in\mathcal{F}_m} \|f\|_{\mathbb{P}_n,2}/\|F_m\|_{\mathbb{P}_n,2} \geq 1/\sqrt{n}$ for $m \leq N$. The latter bound follows from $\mathbb{E}_n[F_m^2] = \mathbb{E}_n[\sup_{f\in\mathcal{F}_m} |f(Z_i)|^2] \leq \sup_{i\leq n} \sup_{f\in\mathcal{F}_m} |f(Z_i)|^2$, and from

$$\sup_{f\in\mathcal{F}_m} \mathbb{E}_n[|f(Z_i)|^2] \geq \sup_{f\in\mathcal{F}_m} \sup_{i\leq n} |f(Z_i)|^2/n. \qquad \square$$

Acknowledgments. We would like to thank Arun Chandrasekhar, Denis Chetverikov, Moshe Cohen, Brigham Fradsen, Joonhwan Lee, Ye Luo and Pierre-Andre Maugis for thorough proofreading of several versions of this paper and their detailed comments that helped us considerably improve the paper. We also would like to thank Don Andrews, Alexandre Tsybakov, the editor Susan Murphy, the Associate Editor and three anonymous referees for their comments that also helped us considerably improve the paper. We would also like to thank the participants of seminars in Brown University, CEMMAP Quantile Regression conference at UCL, Columbia University, Cowles Foundation Lecture at the Econometric Society Summer Meeting, Harvard-MIT, Latin American Meeting 2008 of the Econometric Society, Winter 2007 North American Meeting of the Econometric Society, London Business School, PUC-Rio, the Stats in the Chateau, the Triangle Econometrics Conference and University College London.

SUPPLEMENTARY MATERIAL

Supplement to "ℓ_1-penalized quantile regression in high-dimensional sparse models" (DOI: 10.1214/10-AOS827SUPP; .pdf). We included technical proofs omitted from the main text: Examples of simple sufficient conditions, VC index bounds and Gaussian sparse eigenvalues.

REFERENCES

[1] BELLONI, A. and CHERNOZHUKOV, V. (2009). Computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** 2011–2055. MR2533478

[2] BELLONI, A. and CHERNOZHUKOV, V. (2010). Supplement to "ℓ_1-penalized quantile regression in high-dimensional sparse models." DOI: 10.1214/10-AOS827SUPP.

[3] BELLONI, A. and CHERNOZHUKOV, V. (2009). ℓ_1-penalized quantile regression in high-dimensional sparse models. Available at arXiv:0904.2931.

[4] BELLONI, A. and CHERNOZHUKOV, V. (2008). Conditional quantile processes under increasing dimension. Technical report, Duke and MIT.

[5] BELLONI, A. and CHERNOZHUKOV, V. (2009). Post-ℓ_1-penalized estimators in high-dimensional linear regression models. Available at arXiv:1001.0188.

[6] BERTSIMAS, D. and TSITSIKLIS, J. (1997). *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA.

[7] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[8] BUCHINSKY, M. (1994). Changes in the U.S. wage structure 1963–1987: Application of quantile regression. *Econometrica* **62** 405–458.

[9] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)* (G. Lugosi and H. U. Simon, eds.). *Lecture Notes in Artificial Intelligence* **4005** 379–391. Springer, Berlin. MR2280619

[10] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[11] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[12] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[13] CHERNOZHUKOV, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160

[14] FAN, J. and LV, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[15] GUTENBRUNNER, C. and JUREČKOVÁ, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.* **20** 305–330. MR1150346

[16] HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimenions. *J. Multivariate Anal.* **73** 120–135. MR1766124

[17] KNIGHT, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26** 755–770. MR1626024

[18] KNIGHT, K. and FU, W. J. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[19] KOENKER, R. (2005). *Quantile Regression*. Cambridge Univ. Press, Cambridge. MR2268657

[20] KOENKER, R. (2010). Additive models for quantile regression: Model selection and confidence bandaids. Working paper. Available at http://www.econ.uiuc.edu/~roger/research/bandaids/bandaids.pdf.

[21] KOENKER, R. and BASSET, G. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

[22] KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7–57. MR2500227

[23] LAPLACE, P.-S. (1818). *Théorie Analytique des Probabilités*. Éditions Jacques Gabay (1995), Paris.

[24] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. *Ergebnisse der Mathematik und ihrer Grenzgebiete* **23**. Springer, Berlin. MR1102015

[25] LOUNICI, K., PONTIL, M., TSYBAKOV, A. B. and VAN DE GEER, S. A. (2009). Taking advantage of sparsity in multi-task learning. In *COLT'09*. Omnipress, Madison, WI.

[26] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

[27] PORTNOY, S. (1991). Asymptotic behavior of regression quantiles in nonstationary, dependent cases. *J. Multivariate Anal.* **38** 100–113. MR1128939

[28] PORTNOY, S. and KOENKER, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12** 279–300. MR1619189

[29] ROSENBAUM, M. and TSYBAKOV, A. B. (2010). Sparse recovery under matrix uncertainty. *Ann. Statist.* **38** 2620–2651.

[30] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[31] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge, MA.

[32] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809

[33] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[34] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

FUQUA SCHOOL OF BUSINESS
DUKE UNIVERSITY
1 TOWERVIEW DRIVE
DURHAM, NORTH CAROLINA 27708-0120
PO BOX 90120
USA
E-MAIL: abn5@duke.edu

DEPARTMENT OF ECONOMICS
AND OPERATIONS RESEARCH CENTER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
ROOM E52-262F
CAMBRIDGE, MASSACHUSETTS 02142
USA
E-MAIL: vchern@mit.edu