MIT Open Access Articles

#[subscript 1]-penalized quantile regression in high-dimensional sparse models

Massachusetts Institute of Technology

ℓ_1-PENALIZED QUANTILE REGRESSION IN HIGH-DIMENSIONAL SPARSE MODELS

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 non-zero impact on each conditional quantile of the response variable, where s grows slower 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 for the leading designs 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 for the leading designs 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 $\mathcal{U}$. Finally, we evaluate the performance of ℓ_1-QR and post-ℓ_1-QR in a numerical experiment, and provide an application to testing the validity of the Solow-Swan model for international economic growth.

1. Introduction. Quantile regression is an important statistical method for analyzing the impact of regressors on the conditional distribution of a response variable (cf. [24], [21]). In particular, it captures the heterogeneity of the impact of regressors on the different parts of the distribution [8], exhibits robustness to outliers [20], has excellent computational properties [31], and has wide applicability [20]. The asymptotic theory for quantile regression is well 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 [21], [30], [16], [18], [13] and others. The asymptotic theory under an increasing number of regressors is given in [17] and [3, 5], covering the case where the number of regressors p is negligible relative to the sample size n (i.e., $p = o(n)$).

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 and smaller than the sample size, that is, $s = o(n)$. HDSMs ([7, 12, 29]) 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, primarily focusing on penalized mean regression, with the ℓ_1-norm acting as a penalty function [7, 12, 23, 29, 36, 38]. [7, 12, 23, 29, 38] 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. [36] demonstrated a similar fundamental result on the excess forecasting error loss under both quadratic and non-quadratic 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, 15, 27, 32] 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 not consistent in HDSMs, we consider quantile regression penalized by the ℓ_1-norm of parameter coefficients, denoted ℓ_1-QR. First, we show that ℓ_1-QR estimates of regression coefficients and regression functions are consistent at the near-oracle rate $\sqrt{s/n}\sqrt{\log(p \vee n)}$

in leading designs, uniformly in a compact interval $\mathcal{U} \subset (0, 1)$ of quantile indices.[1] (This result is different and hence complementary to [36]'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 non-zero 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 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 can miss only 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 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 and an international economic growth example. To the best of our knowledge, all of the above results are new and contribute to the literature on HDSMs. We also hope that our results on post-penalized estimators and some proofs could 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 [37]. In particular, given random sample $Z_1, ..., Z_n$, let $\mathbb{G}_n(f) := n^{-1/2} \sum_{i=1}^{n} (f(Z_i) - \mathrm{E}[f(Z_i)])$ and $\mathbb{E}_n f = n^{-1} \sum_{i=1}^{n} f(Z_i)$. We use the notation $a \lesssim b$ to denote $a \leq cb$ 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 non-zero 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 $j \notin T$.

---

[1]Under $s \to \infty$, the oracle rate, uniformly over a proper compact interval $\mathcal{U}$, is $\sqrt{(s/n) \log n}$, cf. [5]; the oracle rate for a single quantile index is $\sqrt{s/n}$, cf. [17].

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 set-up 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 u-th conditional quantile function of y given x is given by

$$(2.1) \qquad Q_{y|x}(u) = x'\beta(u), \quad \beta(u) \in \mathbb{R}^p, \quad u \in \mathcal{U},$$

where $\mathcal{U} \subset (0,1)$ is a compact set of quantile indices. 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

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

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

The population coefficient $\beta(u)$ is known to be a minimizer of the criterion function

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

where $\rho_u(t) = (u - 1\{t \leq 0\})t$ 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):

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

In high-dimensional settings, particularly when $p \geq n$, ordinary quantile regression is generally not consistent, 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 been proven to be quite useful in least squares settings is the ℓ_1-penalty leading to the Lasso estimator [34].

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

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

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 [19] 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

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

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

$$(2.5) \qquad \widetilde{\beta}(u) \in \arg \min_{\beta \in \mathbb{R}^p : \beta_{\widehat{T}_u^c}=0} \widehat{Q}_u(\beta),$$

that is, in (2.5) we remove the regressors that were not selected from further estimation. 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. Specifically, we are interested in the highly realistic scenario where the first step estimator $\widehat{\beta}(u)$ fails to select some the components of $\beta(u)$. Our goal is to derive the rate of convergence for the post-penalized estimator and to show that it can actually 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

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

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 $X = [x_1,\ldots,x_n]'$. Then we set

$$(2.7) \quad \lambda = c \cdot \Lambda(1-\alpha|X), \ \text{where} \ \Lambda(1-\alpha|X) := (1-\alpha)\text{-quantile of } \Lambda \text{ conditional on } X,$$

and the constant $c > 1$ depends on the design.[2] 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 $c = 2$, similarly to [7]'s choice for least squares. Furthermore, we recommend computing $\Lambda(1 - \alpha|X)$ using simulation of Λ.[3]

The parameter $1 - \alpha$ is the confidence level in the sense that our (non-asymptotic) bounds on the estimation error will hold with probability close to $1 - \alpha$. If we want to maximize this confidence level, e.g., as in [7], subject to the estimation error contracting at the optimal rate as $n \to \infty$, then we may set

$$(2.8) \qquad 1 - \alpha = 1 - 1/p \to 1 \text{ as } p \to \infty.$$

This choice of $1 - \alpha$ does not affect the stochastic order of magnitude of $\Lambda(1 - \alpha|X)$ and leads to optimal rates of convergence, as follows from Theorem 1 and 2, respectively. However, a high confidence level has the cost of a high regularization bias. Therefore, if, instead, we want to minimize the regularization bias subject to the estimation error contracting at the optimal rate, then we should set the confidence level $1 - \alpha$ to grow as slowly as possible. As a limit of this rule, we may set

$$(2.9) \qquad 1 - \alpha = 1 - \alpha_0 \text{ for some fixed } 1/p \leq \alpha_0 < 1.$$

In this case, estimation error will contract at the optimal rate with a probability that is close to $1 - \alpha_0$. Our non-asymptotic bounds on the estimation error stated in Theorem 2 expressly allow for any choice of $1 - \alpha$, including either (2.8) or (2.9). Also, in computational experiments, we found that (2.9) with $1 - \alpha_0 = .9$ worked well.

The formal rationale behind the choice (2.7) for the penalty level λ is that this choice precisely leads 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 this general strategy leads precisely to the choice (2.7), because the gradient of the quantile regression objective function evaluated at the truth has a pivotal representation, as we further

---

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

[3]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, whereas the analytical bound effectively treats the columns of X as uncorrelated and is thus more conservative, at least in finite samples.

explain below. This makes our choice of λ independent of the conditional density of y_i given x_i, which is of considerable practical value. In contrast, in the least squares problem, the choice of λ depends on the standard deviation of the regression errors, and also relies on the homoscedasticity and Gaussianity of errors.

A less formal, though more intuitive rationale for the choice (2.7) is as follows. By optimality $\widehat{\beta}(u)$ obeys

$$0 \in \partial \widehat{Q}_u(\widehat{\beta}(u)) + (\lambda/n)\sqrt{u(1-u)} \, \partial\|\widehat{\beta}(u)\|_{1,n} \ \ \text{for all } u \in \mathcal{U},$$

where ∂ is the subdifferential operator. Let $\widehat{S}_u$ denote an element of $\partial \widehat{Q}_u$. Since $\partial\|\beta\|_{1,n} \subseteq \{\widehat{D}v : v \in [-1,1]^p\}$, where $\widehat{D} = \text{diag}[\widehat{\sigma}_1, ..., \widehat{\sigma}_p]$, for each $u \in \mathcal{U}$ there is an element $\widehat{S}_u(\widehat{\beta}(u)) \in \partial \widehat{Q}_u(\widehat{\beta}(u))$ such that $\sup_{u \in \mathcal{U}} n\|\widehat{D}^{-1} \, \widehat{S}_u(\widehat{\beta}(u))/\sqrt{u(1-u)}\|_\infty \leq \lambda$. Then it makes sense to choose λ so that the true value $\beta(u)$ also obeys this constraint with a high probability:

$$(2.10) \qquad P\left\{\Lambda = \sup_{u \in \mathcal{U}} \ n\left\|\widehat{D}^{-1} \, \widehat{S}_u(\beta(u))/\sqrt{u(1-u)}\right\|_\infty \leq \lambda\right\} \geq 1 - \alpha,$$

where $\widehat{S}_u(\beta(u)) = \mathbb{E}_n[(u - 1\{y_i \leq x_i'\beta(u)\})x_i] \in \partial \widehat{Q}_u(\beta(u))$. A key observation is that $\widehat{S}_u(\beta(u)) = \mathbb{E}_n[(u - 1\{u_i \leq u\})x_i]$ 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. *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), and with the first component of x_i equal to one. 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, 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 and x. We assume that $n \wedge p \geq 3$.*

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 commonly made in the literature on HDSMs.

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 $u \in \mathcal{U}$:*

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

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

Condition D.2 imposes sparsity and also 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 $P(\max_{1 \leq j \leq p} |\widehat{\sigma}_j - 1| \leq 1/2) \geq 1 - \gamma \to 1$ as $n \to \infty$.*

Condition D.3 requires that $\widehat{\sigma}_j$ does not deviate too much from σ_j and normalizes $\sigma_j^2 = \mathrm{E}[x_{ij}^2] = 1$. That $1 - \gamma \to 1$ as $n \to \infty$ implicitly restricts the growth of p relative to n; in particular, for bounded or Gaussian regressors this requires $n/\log p \to \infty$. The normalization $\mathrm{E}[x_{ij}^2] = 1$ is not restrictive since we do not exploit it in the construction of the estimator and since we can always rescale the original parameters so that regressors have this property. Indeed, given original parameter $\beta^o(u)$ and regressors x^o, define the rescaled parameter $\beta(u) = D\beta^o(u)$ and $x = D^{-1}x^o$ for $D = \mathrm{diag}[\sigma_1^o, ..., \sigma_n^o]$, $\sigma_j^{o2} = \mathrm{E}[x_{ij}^{o2}]$. Then the rates of convergence in the original parametrization follow from the rates of convergence in the rescaled parametrization from the relation $\|G(\widehat{\beta}^o(u) - \beta^o(u))\|^2 \leq \mathrm{maxeig}(D^{-1}G'GD^{-1})\|\widehat{\beta}(u) - \beta(u)\|^2$ for any matrix G.

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 $\overline{T}_u(\delta, m) \subset \{1, ..., p\} \setminus T_u$ as the support of m largest in absolute value components of the vector δ outside of $T_u = \mathrm{support}(\beta(u))$, 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 matrices $\mathrm{E}[x_i x_i']$ and $J_u = \mathrm{E}[f_{y_i|x_i}(x_i'\beta(u)|x_i) \, x_i x_i']$, $u \in \mathcal{U}$ satisfy*

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

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

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

COMMENT 2.1 (RE condition). The restricted eigenvalue (RE) condition is a quantile analog of [7]'s condition for means. The RE constants κ_m and $\underline{f}$ determine the rate of convergence and can change with n, although in many designs such as Example 1 given below these

constants will be bounded away from zero and from above. [7] prove that the RE condition on the design matrix $\mathrm{E}[x_i x_i']$ is quite general, and is weaker than nearly all other design conditions used in the least squares literature; also, since κ_m is non-increasing in m, $RE(c_0, m)$ for any $m > 0$ implies $RE(c_0, 0)$. The constant $\underline{f}$ controls the modulus of continuity between norms weighted by the design matrix $\mathrm{E}[x_i x_i']$ and the Jacobian matrices J_u. Note that $\underline{f}$ is bounded below by $\underline{f}^o$, the minimal value of the conditional density of y_i evaluated at the conditional quantile $x_i'\beta(u)$:

$$(2.11) \qquad \underline{f} \geq \underline{f}^o := \inf_{u \in \mathcal{U}, x \in \mathrm{support}(x_i)} f_{y_i|x_i}(x'\beta(u)|x),$$

where $\underline{f} = \underline{f}^o$ holds with equality in location models, such as Example 1, but $\underline{f} > \underline{f}^o$ in general. The overall rationale behind the RE condition is that under our choice of penalty level, $\delta = \widehat{\beta}(u) - \beta(u)$ will belong to the restricted set A_u with a high probability, and so identifiability and rates would follow from the behavior of $\delta' J_u \delta / \|\delta\|^2$ characterized by $\underline{f}\kappa_m^2$. Lastly, the additional condition $\log(\underline{f}\kappa_0^2) \lesssim \log(n \vee p)$ requires that $\underline{f}\kappa_0^2$ does not increase faster than some power of $(n \vee p)$. This assumption is mild, since typically we are more concerned with $\underline{f}^{1/2}\kappa_0$ going to zero, and simplifies the statements of the main results.

COMMENT 2.2 (RNI condition). The restricted non-linear 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, in the sense precisely described by Lemma 2. It turns out that this coefficient is well-behaved for several designs of interest. Indeed, if the covariates x have a log-concave density, then

$$q \geq 3\underline{f}^{3/2}/(8K_\ell \bar{f}') \text{ for a universal constant } K_\ell .$$

On the other hand, if the covariates $|x_{ij}|$ are uniformly bounded by K_B for each $j \leq p$, and the $RE(c_0, 0)$ condition holds, then $q \geq 3\underline{f}^{3/2}\kappa_0/(8\bar{f}'K_B(1+c_0)\sqrt{s})$. Indeed, the former bound follows from $\mathrm{E}[|x_i'\delta|^3] \leq K_\ell \mathrm{E}[|x_i'\delta|^2]^{3/2}$ holding for log-concave x for some universal constant K_ℓ by Theorem 5.22 of [28]. The latter bound follows from $\mathrm{E}[|x_i'\delta|^3] \leq \mathrm{E}[|x_i'\delta|^2]K_B\|\delta\|_1 \leq \mathrm{E}[|x_i'\delta|^2]K_B(1+c_0)\sqrt{s}\|\delta_{T_u}\| \leq \mathrm{E}[|x_i'\delta|^2]^{3/2}K_B(1+c_0)\sqrt{s}/\kappa_0$ holding since $\delta \in A_u$ so that $\|\delta\|_1 \leq (1+c_0)\|\delta_{T_u}\|_1 \leq \sqrt{s}(1+c_0)\|\delta_{T_u}\|$.

Finally, we state another condition that is 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 matrices $\mathrm{E}[x_i x_i']$ and J_u, $u \in \mathcal{U}$ satisfy:

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

for some $\widetilde{m} \geq 0$. Moreover,

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

COMMENT 2.3 (SE condition).   We invoke the sparse eigenvalue (SE) condition in order to analyze the post-penalized estimator. This assumption is similar to the conditions used in [29] and [38] to analyze Lasso. Our form of the SE condition is neither less nor more general than the RE condition. The rationale behind this condition is that the post-penalized estimator $\widetilde{\beta}(u)$ will be sparse, of dimension at most $\widehat{s}_u = |\widehat{T}_u| \leq s + \widehat{m}$, where $\widehat{m}$ is the number of unnecessary components, that is, components outside T_u. Therefore, both identifiability and rates of convergence would follow from the behavior of $\delta' J_u \delta / \|\delta\|^2$ characterized by $\widetilde{\underline{f}}_{\widehat{m}} \widetilde{\kappa}_{\widehat{m}}^2$.

COMMENT 2.4 (SNI 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. Similarly to the RNI coefficient, if the covariates x have a log-concave density, then the SNI coefficient satisfies

$$\widetilde{q}_{\widetilde{m}} \geq (3/8) \widetilde{\underline{f}}_{\widetilde{m}}^{3/2} / (K_\ell \bar{f}')$$

and if the covariates $|x_{ij}|$ are uniformly bounded by K_B and SE condition holds, then $\widetilde{q}_{\widetilde{m}} \geq (3/8) \widetilde{\underline{f}}_{\widetilde{m}}^{3/2} \widetilde{\kappa}_{\widetilde{m}} / (\bar{f}' K_B \sqrt{\widetilde{m} + s})$. Note that if the selected model has no unnecessary components ($\widetilde{m} = 0$), condition D.5 is an assumption only on the true support.

EXAMPLE 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, $x_1 = 1$ and

$$x_{-1} \sim N(0, \Sigma), \ \ \Sigma_{ij} = \rho^{|i-j|}, \text{ for a fixed } -1 < \rho < 1.$$

This model implies a linear quantile model with coefficients $\beta_1(u) = \beta_1^o + \sigma \Phi^{-1}(u)$ and $\beta_j(u) = \beta_j^o$ for $j = 2, \ldots, p$, in which conditions D.1-5 are easily met for any compact set of quantile indices $\mathcal{U} \subset (0, 1)$. Indeed, let

$$\bar{f}' = \sup_z \phi'(z/\sigma)/\sigma^2, \ \ \bar{f} = \sup_z \phi(z/\sigma)/\sigma, \ \ \underline{f}^o = \min_{u \in \mathcal{U}} \phi(\Phi^{-1}(u))/\sigma,$$

so that D.1 holds with the constants $\bar{f}$ and $\bar{f}'$. D.2 holds, since $\|\beta(u)\|_0 \leq s \leq \|\beta^o\|_0 + 1$ and $u \mapsto \beta(u)$ is Lipschitz over $\mathcal{U}$ with the constant $L = \sup_{u \in \mathcal{U}} \sigma/\phi(\Phi^{-1}(u))$, which trivially obeys $\log L \lesssim \log(n \vee p)$. D.4 also holds, in particular by Chernoff's tail bound

$$P\left\{\max_{1 \leq j \leq p} |\widehat{\sigma}_j - 1| \leq 1/2\right\} \geq 1 - \gamma = 1 - 2p\exp(-n/24),$$

where $1 - \gamma$ approaches 1 if $n/\log p \to \infty$. Furthermore, the smallest eigenvalue of the population design matrix Σ is at least $(1 - |\rho|)/(2 + 2|\rho|)$ and the maximum eigenvalue is at most $(1 + |\rho|)/(1 - |\rho|)$. Thus, D.4 and D.5 hold with

$$\kappa_m \wedge \widetilde{\kappa}_{\widetilde{m}} \geq \sqrt{(1 - |\rho|)/(2 + 2|\rho|)}, \quad \underline{f} \wedge \underline{\widetilde{f}}_{\widetilde{m}} \geq \underline{f}^o, \quad q \wedge \widetilde{q}_{\widetilde{m}} \geq (3/8)(\underline{f}^{o3/2}/\bar{f}'K_\ell),$$

for all $m, \widetilde{m} \geq 0$ and for a universal constant $K_\ell > 0$ defined in Comment 2.2.

2.5. *Overview of Main Results.* Here we discuss our results under the simple setup of Example 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 non-asymptotic results under general conditions in the subsequent sections. Our first main rate result is that the ℓ_1-QR, with our choice (2.7) of parameter λ, satisfies

$$(2.12) \qquad \sup_{u \in \mathcal{U}} \|\widehat{\beta}(u) - \beta(u)\| \lesssim_p \frac{1}{\underline{f}^{1/2}\kappa_0 \ \underline{f}^{1/2}\kappa_s} \sqrt{\frac{s\log(n \vee p)}{n}},$$

provided that the upper bound on the number of non-zero components s satisfies

$$(2.13) \qquad \frac{\sqrt{s\log(n \vee p)}}{\sqrt{n} \ \underline{f}^{1/2}\kappa_0 q} \to 0.$$

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 (2.13) 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

$$(2.14) \qquad \sup_{u \in \mathcal{U}} \|\widehat{\beta}(u)\|_0 \lesssim_p s.$$

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:

$$(2.15) \qquad T_u = \text{support}(\beta(u)) \subseteq \widehat{T}_u = \text{support}(\widehat{\beta}(u)), \text{ for all } u \in \mathcal{U}.$$

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

$$(2.16) \qquad \sup_{u \in \mathcal{U}} \|\widetilde{\beta}(u) - \beta(u)\| \quad \lesssim_p \quad \frac{1}{\underline{\widetilde{f}}_{\widehat{m}} \widetilde{\kappa}_{\widehat{m}}^2} \sqrt{\frac{\widehat{m} \log(n \vee p) + s \log n}{n}} \quad + $$
$$+ \quad \frac{\sup_{u \in \mathcal{U}} 1\{T_u \nsubseteq \widehat{T}_u\}}{\underline{\widetilde{f}}_{\widehat{m}}^{1/2} \widetilde{\kappa}_{\widehat{m}} \underline{f}^{1/2} \kappa_0} \sqrt{\frac{s \log(n \vee p)}{n}},$$

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 non-zero components s obeys the growth condition (2.13) and

$$(2.17) \qquad \frac{\sqrt{\widehat{m} \log(n \vee p) + s \log n}}{\sqrt{n} \, \underline{\widetilde{f}}_{\widehat{m}}^{1/2} \widetilde{\kappa}_{\widehat{m}} \widetilde{q}_{\widehat{m}}} \to_p 0.$$

We see from (2.16) 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 (2.16) 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 the 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 (2.17) 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, 23, 29, 36, 38]. To the best of our knowledge, our results on post-penalized estimators have 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.[4] Our results on sparsity of ℓ_1-QR and model selection also contribute to the analogous results for mean regression [29]. Also, our rate results for ℓ_1-QR are different from, and hence complementary to, the fundamental results in [36] on the excess forecasting loss under possibly non-quadratic loss functions, which also specializes the results to density estimation, mean regression, and logistic regression. Indeed, in principle we could apply theorems in [36] to the single quantile regression problem to derive the bounds on the excess loss from forecasting y_i with $x_i'\widehat{\beta}(u)$ under loss ρ_u.[5] However, these bounds would not imply our results (2.12), (2.16), (2.14), (2.15), and (2.7), 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 determine the magnitude of our suggested penalty level λ via equation (2.7).

---

[4]In a companion work, we extend our results to least squares and related problems.

[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 (cf. [36]'s conditions I.1 and I.2 and others), and which do not apply in our main examples, e.g., in Example 1 with normal regressors.

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

(i) $P\left(\Lambda \geq k \cdot C_\Lambda \, W_{\mathcal{U}} \sqrt{n \log p} \, |X\right) \leq p^{-k^2+1}$,

(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 $\ell_1 - QR$. We start with the following preliminary result which shows that if the penalty level exceeds the specified threshold, the estimator $\widehat{\beta}(u) - \beta(u)$ will belong to the restricted set A_u.

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

(3.1) $$\frac{2}{3}\|\delta\|_{1,n} \leq \|\delta\|_1 \leq 2\|\delta\|_{1,n}.$$

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

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

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 of [7] for least squares.

LEMMA 2 (Identifiability Relations over Restricted Set). *Condition D.4, namely* $\mathrm{RE}(c_0, m)$ *and* $\mathrm{RNI}(c_0)$, *implies that for any* $\delta \in A_u$ *and* $u \in \mathcal{U}$,

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

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

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

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

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

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 3 (Control of Empirical Error). *Under D.1-4, for any $t > 0$ let*

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

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

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

In order to prove the lemma we use a combination of chaining arguments and exponential inequalities for contractions [25]. Our use of the contraction principle is inspired by its fundamentally innovative use in [36]; 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 1-3, 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 3, D.2, and D.4.

THEOREM 2 (Uniform Bounds on Estimation Error of ℓ_1-QR). *Assume that conditions D.1-4, and let $C > 2C_\Lambda \sqrt{1 + \log(1/\alpha)/\log p} \vee [C_E \sqrt{1 \vee [C_L + C_f + 1/2]}]$. Let λ_0 be defined as in (3.2). Then uniformly in the penalty level λ such that*

$$(3.8) \qquad\qquad \lambda_0 \leq \lambda \leq C \cdot W_{\mathcal{U}} \sqrt{n \log p},$$

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

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

provided s obeys the growth condition

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

This result derives the rate of convergence of the ℓ_1-penalized quantile regression estimator in the intrinsic norm 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. An immediate consequence of this result and of Lemma 2 is the following corollary.

COROLLARY 1. *Under the conditions of Theorem 2, for any $A > 1$ with probability at least $1 - \alpha - 4\gamma - 3p^{-A^2}$,*

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

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

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 non-quadratic 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 3 holds;
(iii) $\Omega_3 :=$ the event in which $\Lambda(1 - \alpha|X) \leq \sqrt{1 + \log(1/\alpha)/\log p} \cdot C_\Lambda W_{\mathcal{U}}\sqrt{n \log p}$.

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

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

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

is impossible, which will prove the theorem. First note that the event in (3.10) implies that for some $u \in \mathcal{U}$

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

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:

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

Also, by inequality (3.4) in Lemma 2, 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

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

Also by Lemma 3, 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

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

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

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

Using the identifiability relation (3.7) stated in Lemma 2, we further get

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

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

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

Note that qt cannot be smaller than $t^2/4$ under the growth condition (3.9) of the theorem. Thus, using also the lower bound on C given in the theorem 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. Therefore, the result follows. $\qquad\square$

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 $\mathbb{E}_n\,[x_i x_i']$, specifically on

$$(3.13) \qquad \varphi(k) = \max_{\|\delta\|=1, \|\delta\|_0 \leq k} \mathrm{E}\left[(x_i'\delta)^2\right] \quad \text{and} \quad \phi(k) = \sup_{\|\delta\|\leq 1, \|\delta\|_0 \leq k} \mathbb{E}_n\left[(x_i'\delta)^2\right] \vee \mathrm{E}\left[(x_i'\delta)^2\right].$$

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

LEMMA 4 (Empirical Pre-Sparsity). *Let $\widehat{s} = \sup_{u \in \mathcal{U}} \|\widehat{\beta}(u)\|_0$. Under D.1-4, 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 non-zero components $\widehat{s}$ as a function of λ and $\phi(\widehat{s})$. Restricting $\lambda \geq 2\sqrt{2}W_{\mathcal{U}}\sqrt{n\log(n \vee p)\phi(n/\log(n \vee p))}$ makes the term $\phi\left(n/\log(n \vee p)\right)$ 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 Example 1, we have $\phi(n/\log(n \vee p)) \lesssim_p 1$ while $\phi(n) \gtrsim_p \sqrt{\log p}$ by [4].

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

LEMMA 5 (Empirical Sparsity). *Assume D.1-4 and let* $r = \sup_{u \in \mathcal{U}} \|J_u^{1/2}(\widehat{\beta}(u) - \beta(u))\|$. *Then, for any* $\varepsilon > 0$, *there is a constant* $K_\varepsilon \geq \sqrt{2}$ *such that with probability at least* $1 - \varepsilon - \gamma$

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

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.1-4 hold, and let* λ *satisfy* $\lambda \geq \lambda_0$ *and*

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

for some constant $K' \geq K \geq 2K_\varepsilon \bar{\phi}_\varepsilon^{1/2}$, *for* K_ε *defined in Lemma 5. Then, for any* $A > 1$ *with probability at least* $1 - \alpha - 2\varepsilon - 4\gamma - p^{-A^2}$

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

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

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

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 Example 1. In particular, (c) follows from the concentration inequalities and from results in classical random matrix theory; see [4] 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 4, $\widehat{s} \leq n/\log(n \vee p)$ with probability $1 - \varepsilon$. With at least the same probability, the choice of λ yields

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

so that by virtue of Lemma 5 and by $\mu(\widehat{s}) \leq \mu := \mu(n/\log(n \vee p))$,

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

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. $\qquad\square$

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*

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

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

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

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*

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

These results parallel analogous results in [29] for mean regression. The first result says that if non-zero 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 zero. The second result states that if the further stated conditions are satisfied, the 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 non-zero 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 the conditions that allow for imperfect model selection, including cases where we miss some non-zero 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 will rely crucially on the identifiability and control of the empirical error over the sparse sets $\widetilde{A}_u(\widetilde{m}) := \{\delta \in \mathbb{R}^p \ : \ \|\delta_{T_u^c}\|_0 \leq \widetilde{m}\}$.

LEMMA 6 (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 $\widetilde{m} \leq n$, we have that*

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

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

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

obeys

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

In order to prove this lemma 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 following theorem establishes the properties of the post-model-selection estimators.

THEOREM 5 (Uniform Bounds on Estimation Error of post-ℓ_1-QR). *Let $\widehat{\beta}(u)$ be any first step estimator, B_n a random variable such that $B_n \geq \sup_{u \in \mathcal{U}} \widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u))$, and $\widetilde{\beta}(u)$ the second step estimator defined as (2.5) for each $u \in \mathcal{U}$. Assume that $|\cup_{u \in \mathcal{U}} T_u| \leq n$, D.1-3 hold, and 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 bound*

$$(3.18) \qquad \sup_{u \in \mathcal{U}} \left\| J_u^{1/2}(\widetilde{\beta}(u) - \beta(u)) \right\| \leq 4A_{\varepsilon,n}/[\underline{\widetilde{f}}_{\widehat{m}}^{1/2}\widetilde{\kappa}_{\widehat{m}}] + \sup_{u \in \mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\}\sqrt{4B_n}$$

holds with probability at least $1 - 2\varepsilon$, where $A_{\varepsilon,n} := C_\varepsilon \sqrt{(\widehat{m}\log(n \vee p) + s\log n)\phi(\widehat{m} + s)/n}$, provided that s obeys the growth condition $4\widetilde{q}_{\widehat{m}}A_{\varepsilon,n}/[\underline{\widetilde{f}}_{\widehat{m}}^{1/2}\widetilde{\kappa}_{\widehat{m}}] + \sup_{u \in \mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\}B_n \leq 4\widetilde{q}_{\widehat{m}}^2$. In particular, when the first step estimator $\widehat{\beta}(u)$ is the ℓ_1-QR estimator defined by (2.4) and D.1-5 hold, the random variable B_n can be defined as

$$(3.19) \qquad B_n = \lambda\sqrt{s}r/(n\underline{f}^{1/2}\kappa_0), \quad r := \sup_{u \in \mathcal{U}} \|J_u^{1/2}(\widehat{\beta}(u) - \beta(u))\|.$$

This theorem describes the performance of a general post-model selection estimator as well as the performance of the post-ℓ_1-QR estimator that results from using ℓ_1-QR as the model selector. After plugging in the bound on the rate r and the choice of λ from Theorem 2 into (3.19), we obtain the following more explicit statement about the performance of post-ℓ_1-QR, as measured by the intrinsic norm.

COROLLARY 2. *Under the conditions of Theorems 2 and 5, when the first step estimator* $\widehat{\beta}(u)$ *is the ℓ_1-QR estimator defined by (2.4), with probability at least* $1 - 3\gamma - 3p^{-A^2} - 2\varepsilon$

$$
\sup_{u\in\mathcal{U}}\left\|J_u^{1/2}(\widetilde{\beta}(u) - \beta(u))\right\| \leq \frac{4C_\varepsilon\sqrt{\phi(\widehat{m}+s)}}{\underline{\widetilde{f}}_{\widehat{m}}^{1/2}\widetilde{\kappa}_{\widehat{m}}} \cdot \sqrt{\frac{\widehat{m}\log(n\vee p) + s\log n}{n}} +
$$
$$
+ \sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} \cdot \frac{4\sqrt{2(1+c_0)A}}{\underline{f}^{1/2}\kappa_0} \cdot C \cdot W_{\mathcal{U}}\sqrt{\frac{s\log(n\vee p)}{n}}.
$$

We can use the following corollary to assess performance in other norms of interest.

COROLLARY 3. *Under conditions of Theorem 5, with probability $1 - \varepsilon$*

$$
\sup_{u\in\mathcal{U}}\left(\mathrm{E}_x[x'(\widetilde{\beta}(u) - \beta(u))]^2\right)^{1/2} \leq \sup_{u\in\mathcal{U}}\left\|J_u^{1/2}(\widetilde{\beta}(u) - \beta(u))\right\|/\underline{\widetilde{f}}_{\widehat{m}}^{1/2},
$$

$$
\sup_{u\in\mathcal{U}}\left\|\widetilde{\beta}(u) - \beta(u)\right\| \leq \sup_{u\in\mathcal{U}}\left\|J_u^{1/2}(\widetilde{\beta}(u) - \beta(u))\right\|/\underline{\widetilde{f}}_{\widehat{m}}^{1/2}\widetilde{\kappa}_{\widehat{m}}.
$$

From Corollaries 2 and 3 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, i.e., $\sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} = 1$ with a non-negligible probability. If (a) $\widehat{m} \leq \widehat{s} \lesssim_p s$, (b) $\phi(\widehat{m}+s) \lesssim_p 1$, and (c) the constants $\underline{\widetilde{f}}_{\widehat{m}}$ 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 Example 1. Recall also that in Example 1 (b) holds by concentration of measure and classical results in random matrix theory, as shown in [4], 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, i.e. $\sup_{u\in\mathcal{U}} 1\{T_u \not\subseteq \widehat{T}_u\} = 0$ 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 of this result, and note that this result could be of interest in other problems.

PROOF OF THEOREM 5. To show the first claim, let

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

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

$$(3.20) \qquad \widehat{Q}_u(\widetilde{\beta}(u)) - \widehat{Q}_u(\beta(u)) \leq 1\{T_u \not\subseteq \widehat{T}_u\}\left(\widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u))\right) \leq 1\{T_u \not\subseteq \widehat{T}_u\}B_n.$$

Also, by Lemma 6, with probability at least $1 - \varepsilon$, we have

$$(3.21) \qquad \sup_{u \in \mathcal{U}} \frac{\epsilon_u(\widetilde{\delta}(u))}{\|\widetilde{\delta}(u)\|} \leq C_\varepsilon \sqrt{\frac{(\widehat{m}\log(n \vee p) + s\log n)\phi(\widehat{m} + s)}{n}} =: A_{\varepsilon,n}.$$

Recall that $\sup_{u \in \mathcal{U}} \|\widetilde{\delta}_{T_u^c}(u)\| \leq \widehat{m} \leq n$ so that by D.5 $t_u \geq \widetilde{\underline{f}}_{\widehat{m}}^{1/2}\widetilde{\kappa}_{\widehat{m}}\|\widetilde{\delta}(u)\|$ for all $u \in \mathcal{U}$ with probability $1 - \varepsilon$. Thus, combining relations (3.20) and (3.21), for every $u \in \mathcal{U}$

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

with probability at least $1 - 2\varepsilon$. Invoking the sparse identifiability relation (3.17) of Lemma 6, with the same probability, for all $u \in \mathcal{U}$,

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

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

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

for every $u \in \mathcal{U}$ and the bounds stated in the theorem now follow.

To show the second claim we note that by the optimality of $\widehat{\beta}(u)$ in (3.16), with probability $1 - \gamma$ we have uniformly in $u \in \mathcal{U}$

$$\widehat{Q}_u(\widehat{\beta}(u)) - \widehat{Q}_u(\beta(u)) \leq \frac{\lambda\sqrt{u(1-u)}}{n}(\|\beta(u)\|_{1,n} - \|\widehat{\beta}(u)\|_{1,n})$$

$$(3.22)$$

$$\leq \frac{\lambda\sqrt{u(1-u)}}{n}\|\widehat{\delta}_{T_u}(u)\|_{1,n} \leq \frac{\lambda\sqrt{u(1-u)}}{n}2\|\widehat{\delta}_{T_u}(u)\|_1,$$

where the last term in (3.22) is bounded by

$$(3.23) \qquad \frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}\|J_u^{1/2}\widehat{\delta}(u)\|}{\underline{f}^{1/2}\kappa_0} \leq \frac{\lambda\sqrt{u(1-u)}}{n}\frac{2\sqrt{s}r}{\underline{f}^{1/2}\kappa_0} \leq B_n,$$

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

4. Empirical Performance. In order to access the finite sample practical performance of the proposed estimators, we conducted a Monte Carlo study and an application to international economic growth.

4.1. *Monte Carlo Simulation.* In this section we will compare the performance of the canonical, ℓ_1-penalized, post-ℓ_1-penalized, and the ideal oracle quantile regression estimators. Recall that the post-penalized estimator applies the canonical quantile regression to the model selected by the penalized estimator. The oracle estimator applies the canonical quantile regression on 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 standard deviations of these estimators.

We begin by considering the following regression model, as in Example 1, where

$$y = x'\beta(0.5) + \varepsilon, \quad \beta(0.5) = (1, 1, 1, 1, 1, 0, \ldots, 0)',$$

where x consists of an intercept and covariates $x_{-1} \sim N(0, \Sigma)$, and the errors ε are independently and identically distributed $\varepsilon \sim N(0, 1)$. We set the dimension p of covariates x equal to 1000, and the dimension s of the true model to 5, and the sample size n to 200. We set the regularization parameter λ equal to the 0.9-quantile of the pivotal random variable Λ, following our proposal in Section 2. We consider two variants of the model above with uncorrelated and correlated regressors, namely $\rho = 0$ and $\rho = 0.5$. We summarize the results on model selection performance of the penalized estimator in Figures 1-2. In the left panels of Figures 1-2, we plot the frequencies of the dimensions of the selected model; in the right panel we plot the frequencies of selecting the correct components. From the right panels we see that the model selection performance is particularly good. From the left panels we see that the frequency of selecting a much larger model than the true model is very small. We also see that in the design with correlated regressors, the performance of the estimator is quite good, as we would expect from our theoretical results. These results confirm the theoretical results of Theorem 4, namely, that when the non-zero coefficients are well-separated from zero, with probability tending to one, the penalized estimator should select the 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 the estimation performance in Table 1. We see that the penalized quantile regression estimator significantly outperforms the canonical quantile regression, as

we would expect from Theorem 2 and from inconsistency of the latter when the number of regressors is larger than the sample size. The 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. Furthermore, we see that the post-penalized quantile regression drastically improves upon the penalized quantile regression, particularly in terms of drastically reducing the bias. The post-penalized estimator in fact does almost as well as the ideal oracle estimator. We also see that the (unarbitrary) correlation of regressors does not harm the performance of the penalized and the two-step estimators, which we would expect from our theoretical results. In summary, we find the estimation performance of the penalized and two-step estimators to be in agreement with our theoretical results.

MONTE CARLO RESULTS

Isotropic Gaussian Design

	Mean ℓ_0 norm	Mean ℓ_1 norm	Bias	Std Deviation
Canonical QR	1000	25.27	1.6929	0.99
Penalized QR	5.14	2.43	1.1519	0.37
Post-Penalized QR	5.14	4.97	0.0276	0.29
Oracle QR	5.00	5.00	0.0012	0.20

Correlated Gaussian Design

	Mean ℓ_0 norm	Mean ℓ_1 norm	Bias	Std Deviation
Canonical QR	1000	29.40	1.2526	1.11
Penalized QR	5.19	4.09	0.4316	0.29
Post-Penalized QR	5.19	5.02	0.0075	0.27
Oracle QR	5.00	5.00	0.0013	0.25

TABLE 1

The table displays the average ℓ_0 and ℓ_1 norm of the estimators as well as mean bias and standard deviation. We obtained the results using 5000 Monte Carlo repetitions for each design.

4.2. *International Economic Growth Example.* In this section we apply ℓ_1-penalized quantile regression to an international economic growth example, using it primarily as a method for model selection. We use the Barro and Lee data consisting of a panel of 138 countries for the period of 1960 to 1985. We consider the national growth rates in gross domestic product (GDP) per capita as a dependent variable y for the periods 1965-75 and 1975-85.[6] In our analysis, we will consider a model with $p = 60$ covariates, which allows for a total of $n = 90$ complete observations. Our goal here is to select a subset of these covariates and briefly

---

[6]The growth rate in GDP over a period from t_1 to t_2 is commonly defined as $\log(GDP_{t_2}/GDP_{t_1}) - 1$.
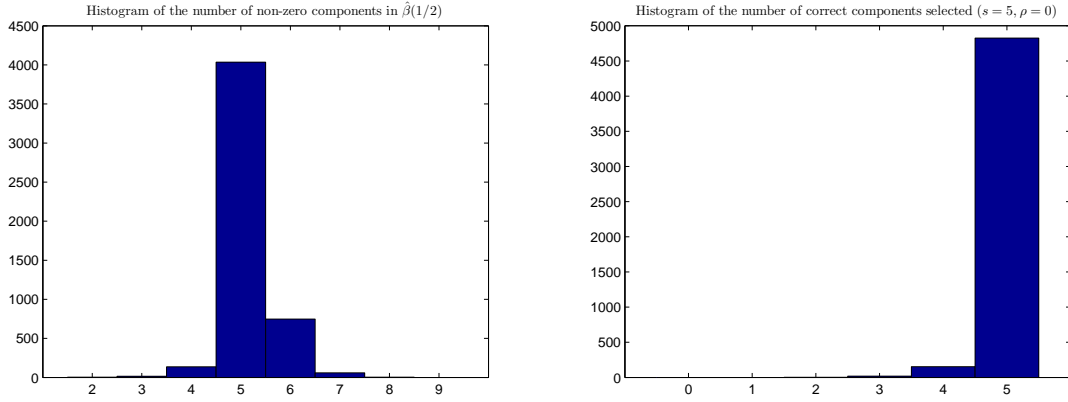
FIG 1. *The figure summarizes the covariate selection results for the isotropic normal design example, based on* 5000 *Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible 1000 covariates. The right panel plots the histogram for the number of significant covariates selected; there are in total* 5 *significant covariates amongst* 1000 *covariates.*
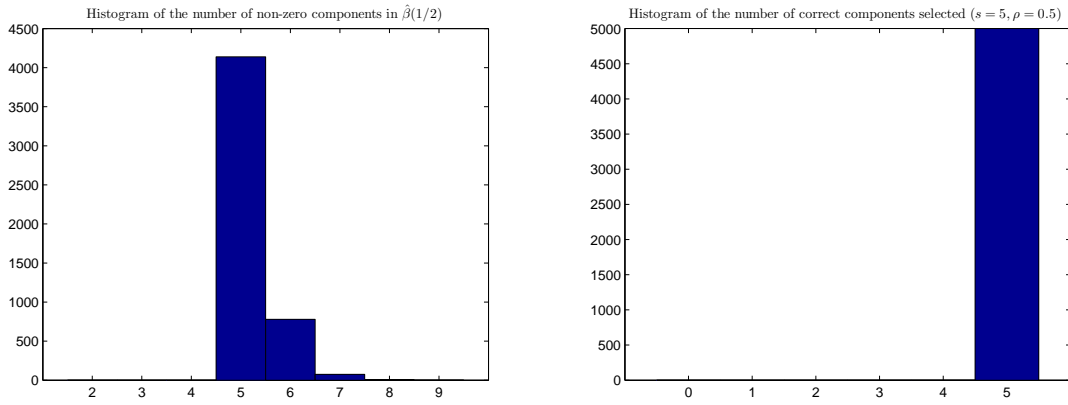

FIG 2. *The figure summarizes the covariate selection results for the correlated normal design example with correlation coefficient* $\rho = .5$*, based on* 5000 *Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible 1000 covariates. The right panel plots the histogram for the number of significant covariates selected; there are in total* 5 *significant covariates amongst* 1000 *covariates. We obtained the results using* 5000 *Monte Carlo repetitions.*

compare the resulting models to the standard models used in the empirical growth literature (Barro and Sala-i-Martin [1], Koenker and Machado [22]).

One of the central issues in the empirical growth literature is the estimation of the effect of an initial (lagged) level of GDP per capita on the growth rates of GDP per capita. In particular, a key prediction from the classical Solow-Swan-Ramsey growth model is the hypothesis of convergence, which states that poorer countries should typically grow faster and therefore should tend to catch up with the richer countries. Thus, such a hypothesis states that the effect of the initial level of GDP on the growth rate should be negative. As pointed out in Barro and Sala-i-Martin [2], this hypothesis is rejected using a simple bivariate regression of growth rates on the initial level of GDP. (In our case, median regression yields a positive coefficient of 0.00045.) In order to reconcile the data and the theory, the literature has focused on estimating the effect *conditional* on the pertinent characteristics of countries. Covariates that describe such characteristics can include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others [2]. The theory then predicts that for countries with similar other characteristics the effect of the initial level of GDP on the growth rate should be negative ([2])

Given that the number of covariates we can condition on is comparable to the sample size, covariate selection becomes an important issue in this analysis ([26], [33]). In particular, previous findings came under severe criticism for relying on ad hoc procedures for covariate selection. In fact, in some cases, all of the previous findings have been questioned ([26]). Since the number of covariates is high, there is no simple way to resolve the model selection problem using only classical tools. Indeed the number of possible lower-dimensional models is very large, although [26] and [33] attempt to search over several millions of these models. Here we use the Lasso selection device, specifically ℓ_1-penalized median regression, to resolve this important issue.

Let us now turn to our empirical results. We performed covariate selection using ℓ_1-penalized median regression, where we initially used our data-driven choice of penalization parameter λ. This initial choice led us to select no covariates, which is consistent with the situations in which the true coefficients are not well-separated from zero. We then proceeded to slowly decrease the penalization parameter in order to allow for some covariates to be selected. We present the model selection results in Table 3. With the first relaxation of the choice of λ, we select the black market exchange rate premium (characterizing trade openness) and a measure of political instability. With a second relaxation of the choice of λ we select an additional set of educational attainment variables, and several others reported in

the table. With a third relaxation of λ we include yet another set of variables also reported in the table. We refer the reader to [1] and [2] for a complete definition and discussion of each of these variables.

We then proceeded to apply ordinary median regression to the selected models and we also report the standard confidence intervals for these estimates. Table 2 shows these results. We should note that the confidence intervals do not take into account that we have selected the models using the data. (In an ongoing companion work, we are working on devising procedures that will account for this.) We find that in all models with additional selected covariates, the median regression coefficients on the initial level of GDP is always negative and the standard confidence intervals do not include zero. Similar conclusions also hold for quantile regressions with quantile indices in the middle range. In summary, we believe that our empirical findings support the hypothesis of convergence from the classical Solow-Swan-Ramsey growth model. Of course, it would be good to find formal inferential methods to fully support this hypothesis. Finally, our findings also agree and thus support the previous findings reported in Barro and Sala-i-Martin [1] and Koenker and Machado [22].

CONFIDENCE INTERVALS AFTER MODEL SELECTION FOR THE INTERNATIONAL GROWTH REGRESSIONS

Penalization Parameter	Real GDP per capita (log)	
$\lambda = 1.077968$	Coefficient	90% Confidence Interval
$\lambda/2$	-0.01691	$[-0.02552, -0.00444]$
$\lambda/3$	-0.04121	$[-0.05485, -0.02976]$
$\lambda/4$	-0.04466	$[-0.06510, -0.03410]$
$\lambda/5$	-0.05148	$[-0.06521, -0.03296]$

TABLE 2

The table above displays the coefficient and a 90% confidence interval associated with each model selected by the corresponding penalty parameter. The selected models are displayed in Table 3.

APPENDIX A: PROOF OF THEOREM 1

PROOF OF THEOREM 1. We note $\Lambda \leq W_{\mathcal{U}} \max_{1 \leq j \leq p} \sup_{u \in \mathcal{U}} n\mathbb{E}_n\left[(u - 1\{u_i \leq u\})x_{ij}/\widehat{\sigma}_j\right]$. For any $u \in \mathcal{U}$, $j \in \{1, \ldots, p\}$ we have by Lemma 1.5 in [25] 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 [37], with $\widetilde{K} \geq 2\sqrt{\log 2}$ we have

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

MODEL SELECTION RESULTS FOR THE INTERNATIONAL GROWTH REGRESSIONS

Penalization Parameter $\lambda = 1.077968$	Real GDP per capita (log) is included in all models Additional Selected Variables
λ	-
$\lambda/2$	Black Market Premium (log) Political Instability
$\lambda/3$	Black Market Premium (log) Political Instability Measure of tariff restriction Infant mortality rate Ratio of real government "consumption" net of defense and education Exchange rate % of "higher school complete" in female population % of "secondary school complete" in male population
$\lambda/4$	Black Market Premium (log) Political Instability Measure of tariff restriction Infant mortality rate Ratio of real government "consumption" net of defense and education Exchange rate % of "higher school complete" in female population % of "secondary school complete" in male population Female gross enrollment ratio for higher education % of "no education" in the male population Population proportion over 65 Average years of secondary schooling in the male population
$\lambda/5$	Black Market Premium (log) Political Instability Measure of tariff restriction Infant mortality rate Ratio of real government "consumption" net of defense and education Exchange rate % of "higher school complete" in female population % of "secondary school complete" in male population Female gross enrollment ratio for higher education % of "no education" in the male population Population proportion over 65 Average years of secondary schooling in the male population Growth rate of population % of "higher school attained" in male population Ratio of nominal government expenditure on defense to nominal GDP Ratio of import to GDP

TABLE 3

For this particular decreasing sequence of penalization parameters we obtained nested models.

where $\mathbb{G}_n^o$ denotes the symmetrized empirical process (see [37]) generated by the Rademacher variables $\varepsilon_i, i = 1, ..., n$, which are independent of $U = (u_1, ..., u_n)$ and $X = (x_1, ..., x_n)$. 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 [37] 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 [37].

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, e.g., [37]. Since $\|F_j\|_{\mathbb{P}_n,2} \leq 1$ and $\rho(\mathcal{F}_j, \mathbb{P}_n) \leq 1$, we have

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

By Lemma 14 with $D = 1$, there is a universal constant c such that for any $K \geq 1$:

$$P\left(\sup_{f \in \mathcal{F}_j} |\mathbb{G}_n^o(f)| > Kc\bar{e}|X,U\right) \leq \int_0^{1/2} \varepsilon^{-1} n(\varepsilon, \mathcal{F}_j)^{-(K^2-1)} d\varepsilon$$

$$\text{(A.2)} \qquad\qquad\qquad\qquad \leq (1/2)[6C_1'(16e)^6]^{-(K^2-1)} \frac{(1/2)^{10(K^2-1)}}{10(K^2-1)}.$$

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

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

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 side of the display above equal to α. $\qquad\square$

APPENDIX B: PROOFS OF LEMMAS 1-3 (USED IN THEOREM 2)

PROOF OF LEMMA 1. (Restricted Set) Part 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).

Part 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

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

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

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

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

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

Furthermore, since $\left|\widehat{\beta}_j(u) - \beta_j(u)\right| + |\beta_j(u)| - |\widehat{\beta}_j(u)| = 0$ if $\beta_j(u) = 0$, i.e. $j \in T_u^c$,

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

(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 7, $\|\widehat{\beta}(u)\|_0 \leq n$ with probability 1 uniformly in $u \in \mathcal{U}$. $\qquad\square$

PROOF OF LEMMA 2. (Identification in Population) Part 1. Proof of claims (3.3)-(3.5). By $\mathrm{RE}(c_0, m)$ and by $\delta \in A_u$

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

Part 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 $|T_u| \leq s$

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

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

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

Part 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

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

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

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

Step 2. $(r_{A_u} \geq 4q)$ Let $F_{y|x}$ denote the conditional distribution of y given x. From [18], for any two scalars w and v we have that

$$(\text{B.3}) \qquad \rho_u(w - v) - \rho_u(w) = -v(u - 1\{w \leq 0\}) + \int_0^v (1\{w \leq z\} - 1\{w \leq 0\})dz.$$

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

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

Note that for $\delta \in A_u$, if $\|J_u^{1/2}\delta\| \leq 4q \leq (3/2) \cdot (\underline{f}^{3/2}/\bar{f}') \cdot \inf_{\delta \in A_u, \delta \neq 0} \mathrm{E}\left[|x'\delta|^2\right]^{3/2} / \mathrm{E}\left[|x'\delta|^3\right]$, 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 $r_{A_u} \geq 4q$. $\qquad \square$

PROOF OF LEMMA 3. (Control of Empirical Error) We divide the proof in four steps.

Step 1. (Main Argument) Let

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

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

In order to apply symmetrization lemma, Lemma 2.3.7 in [37], 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

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

Then application of the symmetrization lemma for probabilities, Lemma 2.3.7 in [37], yields

$$(B.5) \qquad P(\mathcal{A}(t) \geq M) \leq \frac{2P(\mathcal{A}^o(t) \geq M/4)}{1 - t^2/(\underline{f}M^2)} \leq \frac{2P(\mathcal{A}^o(t) \geq M/4|\Omega_1) + 2P(\Omega_1^c)}{1 - t^2/(\underline{f}M^2)},$$

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 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

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

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 Lemma 1 and 2 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 $u \in \mathcal{U}$

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

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

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

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

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

Then we compute the bounds

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

so that

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

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

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

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

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

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

holding for symmetric random variables z_j, and the last inequality follows from the law of iterated expectations and from $\mathrm{E}[\exp(2\lambda t\Gamma\mathbb{G}_n^o(x_{ij})/\widehat{\sigma}_j)|\Omega_1, X] \leq \exp((2\lambda t\Gamma)^2/2)$ holding by the Hoeffding inequality (more precisely, by the intermediate step in the proof of the Hoeffding inequality, see, e.g., p. 100 in [37]). 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

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

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

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

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

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

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

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 [25]) 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. $\square$

APPENDIX C: PROOF OF LEMMAS 4-5 (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:

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

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

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

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 7 (Signs and Interpolation Property). *(1) For any $j \in \{1,\ldots,p\}$*

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

(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 7. Step 1. Part (1) follows from the complementary slackness condition for linear programming problems, see Theorem 4.5 of [6].

Step 2. To show part (2) consider any $u \in \mathcal{U}$. Trivially we have $\|\widehat{\beta}(u)\|_0 \leq p$. Let $Y = (y_1, \ldots, y_n)'$, $\widehat{\sigma} = (\widehat{\sigma}_1, \ldots, \widehat{\sigma}_p)'$, X be the $n \times p$ matrix with rows $x_i', i = 1, \ldots, n$, $c_u = (ue', (1-u)e', \lambda\sqrt{u(1-u)}\widehat{\sigma}', \lambda\sqrt{u(1-u)}\widehat{\sigma}')'$, and $A = [I \ {-I} \ X \ {-X}]$, where $e = (1, 1, \ldots, 1)'$ denotes an n-vectors of ones, and I denotes the $n \times n$ identity matrix. For $w = (\xi^+, \xi^-, \beta^+, \beta^-)$, the primal problem (C.1) can be written as $\min_w \{c_u'w \ : \ Aw = Y, w \geq 0\}$. Matrix A has rank n, since it has linearly independent rows. By Theorem 2.4 of [6] there is at least one optimal basic solution $\widehat{w}(u) = (\widehat{\xi}^+(u), \widehat{\xi}^-(u), \widehat{\beta}^+(u), \widehat{\beta}^-(u))$, and all basic solutions have at most n non-zero components. Since $\widehat{\beta}(u) = \widehat{\beta}^+(u) - \widehat{\beta}^-(u)$, $\widehat{\beta}(u)$ has at most n non-zero components.

Let I_u denote the number of interpolated points in (2.4) at the quantile index u. We have that $n - I_u$ components of $\widehat{\xi}^+(u)$ and $\widehat{\xi}^-(u)$ are non-zero. Therefore, $\|\widehat{\beta}(u)\|_0 + (n - I_u) \leq n$, which leads to $\|\widehat{\beta}(u)\|_0 \leq I_u$. By step 3 below this holds with equality with probability 1 uniformly over $u \in \mathcal{U}$, thus establishing part (3).

Step 3. Consider the dual problem $\max_a \{Y'a : A'a \leq c_u\}$ for all $u \in \mathcal{U}$. Conditional on X the feasible region of this problem is the polytope $R_u = \{a : A'a \leq c_u\}$. Since $c_u > 0$, R_u is non-empty for all $u \in \mathcal{U}$. Moreover, the form of A' implies that $R_u \subset [-1, 1]^n$ so R_u is bounded. Therefore, if the solution of the dual is not unique for some $u \in \mathcal{U}$ there exist vertices a^1, a^2 connected by an edge of R_u such that $Y'(a^1 - a^2) = 0$. Note that the matrix A' is the same for all $u \in \mathcal{U}$ so that the direction $\frac{a^1 - a^2}{\|a^1 - a^2\|}$ of the edge linking a^1 and a^2 is generated by a finite number of intersections of hyperplanes associated with the rows of A'. Thus, the event $Y'(a^1 - a^2) = 0$ is a zero probability event uniformly in $u \in \mathcal{U}$ since Y is absolutely continuous conditional on X and the number of different edge directions is finite. Therefore the dual problem has a unique solution with probability one uniformly in $u \in \mathcal{U}$. If the dual basic solution is unique, we have that the primal basic solution is non-degenerate, that is, the number of non-zero variables equals n, see [6]. Therefore, with probability one $\|\widehat{\beta}(u)\|_0 + (n - I_u) = n$, or $\|\widehat{\beta}(u)\|_0 = I_u$ for all $u \in \mathcal{U}$. □

PROOF OF LEMMA 4. (Empirical Pre-Sparsity) That $\widehat{s} \leq n \wedge p$ follows from Lemma 7. 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

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

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

To establish the second result, note that $\widehat{s} \leq \bar{m} = \max \{ m : m \leq n \wedge p \wedge 4n^2 \phi(m) W_{\mathcal{U}}^2 / \lambda^2 \}$. Suppose that $\bar{m} > m_0 = n / \log(n \vee p)$, so that $\bar{m} = m_0 \ell$ for some $\ell > 1$, since $\bar{m} \leq n$ is finite. By definition, $\bar{m}$ satisfies $\bar{m} \leq 4n^2 \phi(\bar{m}) W_{\mathcal{U}}^2 / \lambda^2$. Insert the lower bound on λ, m_0, and $\bar{m} = m_0 \ell$ in this inequality, and using Lemma 11 we obtain:

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

which is a contradiction. $\square$

PROOF OF LEMMA 5. (Empirical Sparsity) It is convenient to define:

1. the true rank scores, $a_i^*(u) = u - 1\{y_i \leq x_i' \beta(u)\}$ for $i = 1, \ldots, n$;
2. the estimated rank scores, $a_i(u) = u - 1\{y_i \leq x_i' \widehat{\beta}(u)\}$ for $i = 1, \ldots, n$;
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 $\tilde{x}_{i\widehat{T}_u} = (x_{ij}/\widehat{\sigma}_j, j \in \widehat{T}_u)'$, and $\widehat{\beta}_{\widehat{T}_u}(u) = (\widehat{\beta}_j(u), j \in \widehat{T}_u)'$. From the complementary slackness characterizations (C.3)

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

Therefore we can bound the number $\widehat{s}_u$ of non-zero 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

$$
\lambda \sqrt{\widehat{s}} \leq \sup_{u \in \mathcal{U}} \left\{ \frac{\left\| n \mathbb{E}_n \left[\tilde{x}_{i\widehat{T}_u} (\widehat{a}_i(u) - a_i(u)) \right] \right\| + \left\| n \mathbb{E}_n \left[\tilde{x}_{i\widehat{T}_u} (a_i(u) - a_i^*(u)) \right] \right\| + \left\| n \mathbb{E}_n \left[\tilde{x}_{i\widehat{T}_u} a_i^*(u) \right] \right\|}{\sqrt{u(1-u)}} \right\}.
$$

This leads to the inequality

$$\lambda\sqrt{\widehat{s}} \;\leq\; \frac{W_{\mathcal{U}}}{\min\limits_{j=1,\ldots,p}\widehat{\sigma}_j}\left(\sup_{u\in\mathcal{U}}\left\|n\mathbb{E}_n\left[x_{i\widehat{T}_u}\left(\widehat{a}_i(u)-a_i(u)\right)\right]\right\| + \sup_{u\in\mathcal{U}}\left\|n\mathbb{E}_n\left[x_{i\widehat{T}_u}\left(a_i(u)-a_i^*(u)\right)\right]\right\|\right) + \\ + \sup_{u\in\mathcal{U}}\left\|n\mathbb{E}_n\left[\tilde{x}_{i\widehat{T}_u}a_i^*(u)/\sqrt{u(1-u)}\right]\right\|.$$

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 7 the penalized quantile regression fit can interpolate at most $\widehat{s}_u \leq \widehat{s}$ points with probability one uniformly over $u\in\mathcal{U}$. This implies that $\mathbb{E}_n\left[|\widehat{a}_i(u)-a_i(u)|^2\right]\leq \widehat{s}/n$. Therefore,

$$\sup_{u\in\mathcal{U}}\left\|n\mathbb{E}_n\left[x_{i\widehat{T}_u}\left(\widehat{a}_i(u)-a_i(u)\right)\right]\right\| \leq n\sup_{\|\alpha\|_0\leq\widehat{s},\|\alpha\|\leq 1}\sup_{u\in\mathcal{U}}\mathbb{E}_n\left[|\alpha'x_i|\,|\widehat{a}_i(u)-a_i(u)|\right]$$
$$\leq n\sup_{\|\alpha\|_0\leq\widehat{s},\|\alpha\|\leq 1}\sqrt{\mathbb{E}_n\left[|\alpha'x_i|^2\right]}\sup_{u\in\mathcal{U}}\sqrt{\mathbb{E}_n\left[|\widehat{a}_i(u)-a_i(u)|^2\right]} \leq \sqrt{n\phi(\widehat{s})\widehat{s}}.$$

(b) To bound the second term, note that

$$\sup_{u\in\mathcal{U}}\left\|n\mathbb{E}_n\left[x_{i\widehat{T}_u}\left(a_i(u)-a_i^*(u)\right)\right]\right\|$$
$$\leq \sup_{u\in\mathcal{U}}\left\|\sqrt{n}\,\mathbb{G}_n\left(x_{i\widehat{T}_u}\left(a_i(u)-a_i^*(u)\right)\right)\right\| + \sup_{u\in\mathcal{U}}\left\|n\mathbb{E}\left[x_{i\widehat{T}_u}\left(a_i(u)-a_i^*(u)\right)\right]\right\|$$
$$\leq \sqrt{n}\epsilon_1(r,\widehat{s}) + \sqrt{n}\epsilon_2(r,\widehat{s}).$$

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

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

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

By Lemma 10 there is a constant $A_{\varepsilon/2}^1$ such that $\sqrt{n}\epsilon_1(r,\widehat{s})\leq A_{\varepsilon/2}^1\sqrt{n\widehat{s}\log(n\vee p)}\sqrt{\phi(\widehat{s})}$ with probability $1-\varepsilon/2$. By Lemma 8 we have $\sqrt{n}\epsilon_2(r,\widehat{s})\leq n(\mu(\widehat{s})/2)(r\wedge 1)$.

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

$$\sup_{u\in\mathcal{U}}\left\|n\mathbb{E}_n\left[\tilde{x}_{i\widehat{T}_u}a_i^*(u)/\sqrt{u(1-u)}\right]\right\| \leq \sqrt{\widehat{s}}\Lambda \leq \sqrt{\widehat{s}}\,A_{\varepsilon/2}^0\,W_{\mathcal{U}}\sqrt{n\log p},$$

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

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

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

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

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

LEMMA 8 (Controlling linearization error ϵ_2). *Under D.1-2*

$$\epsilon_2(r, m) \leq \sqrt{n}\sqrt{\varphi(m)} \left\{ 1 \wedge \left(2[\bar{f}/\underline{f}^{1/2}]r \right) \right\} \; \text{for all } r > 0 \text{ and } m \leq n.$$

PROOF. By definition

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

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

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

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

$$\mathrm{E}[(1\{y_i \leq x_i'\beta\} - 1\{y_i < x_i'\beta(u)\})^2] \leq \mathrm{E}\left[1\{|y_i - x_i'\beta(u)| \leq |x_i'(\beta - \beta(u))|\}\right]$$
$$\leq \mathrm{E}\left[(2\bar{f}|x_i'(\beta - \beta(u))|) \wedge 1\right] \leq \left\{2\bar{f}\left(\mathrm{E}[|x_i'(\beta - \beta(u))|^2]\right)^{1/2}\right\} \wedge 1$$

and $\left(\mathrm{E}[|x_i'(\beta - \beta(u))|^2]\right)^{1/2} \leq \|J_u^{1/2}(\beta - \beta(u))\|/\underline{f}^{1/2}$ by Lemma 2, the result follows. $\qquad \square$

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 ([37]) of a relevant function class.

LEMMA 9. *(1) Consider a fixed subset $T \subset \{1, 2, \ldots, p\}$, $|T| = m$. The class of functions*

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

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

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

PROOF. The proof of part (1) follows by showing that the corresponding subgraph class is created by at most K operations of taking unions, intersections, and complements of VC classes of sets with VC index at most m, and then appealing to [37] Lemma 2.6.17. We relegate the details to [4] for brevity.

To show part (2) let $\mathcal{F}_T$ denote a restriction of $\mathcal{F}_m$ for a particular choice of m non-zero components. Part (1) implies $N(\epsilon \|F_T\|_{Q,2}, \mathcal{F}_T, L_2(Q)) \leq C(cm)(16e)^{cm}(1/\epsilon)^{2(cm-1)}$, where C is a universal constant (see [37] Theorem 2.6.7). Since we have at most $\binom{p}{m} \leq (ep/m)^m$ different restrictions T, the total covering number is bounded according the statement of the lemma. □

LEMMA 10 (Controlling empirical error ϵ_1). *Under D.1-2 there exists a universal constant A such that with probability $1 - \delta$*

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

PROOF. By definition $\epsilon_1(r, m) \leq \sup_{f \in \mathcal{F}_m} |\mathbb{G}_n(f)|$. From Lemma 9 the uniform covering number of $\mathcal{F}_m$ is bounded by $C (16e/\epsilon)^{2(cm-1)} (ep/m)^m$. Using Lemma 17 with $\theta_m = p$ we have that uniformly in $m \leq n$, with probability at least $1 - \delta$

(C.7) $$\sup_{f \in \mathcal{F}_m} |\mathbb{G}_n(f)| \leq A\delta^{-1/2} \sqrt{m \log(n \vee p)} \max \left\{ \sup_{f \in \mathcal{F}_m} \mathrm{E}[f^2]^{1/2}, \sup_{f \in \mathcal{F}_m} \mathbb{E}_n[f^2]^{1/2} \right\}$$

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

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

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

PROOF. Let $\bar{\alpha}$ achieve $\phi_M(\ell k)$. Moreover let $\sum_{i=1}^{\lceil \ell \rceil} \alpha_i = \bar{\alpha}$ such that $\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|_0 = \|\bar{\alpha}\|_0$. We can choose α_i's such that $\|\alpha_i\|_0 \leq k$ since $\lceil \ell \rceil k \geq \ell k$. Since M is positive semi-definite, for any i, j w $\alpha_i' M \alpha_i + \alpha_j' M \alpha_j \geq 2 |\alpha_i' M \alpha_j|$. Therefore

$$
\begin{aligned}
\phi_M(\ell k) &= \bar{\alpha}' M \bar{\alpha} = \sum_{i=1}^{\lceil \ell \rceil} \alpha_i' M \alpha_i + \sum_{i=1}^{\lceil \ell \rceil} \sum_{j \neq i} \alpha_i' M \alpha_j \leq \sum_{i=1}^{\lceil \ell \rceil} \{\alpha_i' M \alpha_i + (\lceil \ell \rceil - 1)\alpha_i' M \alpha_i\} \\
&\leq \lceil \ell \rceil \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 \phi_M(\|\alpha_i\|_0) \leq \lceil \ell \rceil \max_{i=1,\ldots,\lceil \ell \rceil} \phi_M(\|\alpha_i\|_0) \leq \lceil \ell \rceil \phi_M(k)
\end{aligned}
$$

where we used that $\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 = 1$. $\qquad \square$

APPENDIX D: PROOF OF THEOREM 4

PROOF OF THEOREM 4. By assumption $\sup_{u \in \mathcal{U}} \|\widehat{\beta}(u) - \beta(u)\|_\infty \leq \sup_{u \in \mathcal{U}} \|\widehat{\beta}(u) - \beta(u)\| \leq r^o < \inf_{u \in \mathcal{U}} \min_{j \in T_u} |\beta_j(u)|$, which immediately implies the inclusion event (3.15), since the converse of this event implies $\|\widehat{\beta}(u) - \beta(u)\|_\infty \geq \inf_{u \in \mathcal{U}} \min_{j \in T_u} |\beta_j(u)|$.

Consider the hard-thresholded estimator next. To establish the inclusion, we note that $\inf_{u \in \mathcal{U}} \min_{j \in T_u} |\widehat{\beta}_j(u)| \geq \inf_{u \in \mathcal{U}} \min_{j \in T_u} \{|\beta_j(u)| - |\beta_j(u) - \widehat{\beta}_j(u)|\} > \inf_{u \in \mathcal{U}} \min_{j \in T_u} |\beta_j(u)| - r^o > \gamma$, by assumption on γ. Therefore $\inf_{u \in \mathcal{U}} \min_{j \in T_u} |\widehat{\beta}_j(u)| > \gamma$ and support $(\beta(u)) \subseteq$ support $(\bar{\beta}(u))$ for all $u \in \mathcal{U}$. To establish the opposite inclusion, consider $e_n = \sup_{u \in \mathcal{U}} \max_{j \notin T_u} |\widehat{\beta}_j(u)|$. By definition of r^o, $e_n \leq r^o$ and therefore $e_n < \gamma$ by the assumption on γ. By the hard-threshold rule, all components smaller than γ are excluded from the support of $\bar{\beta}(u)$ which yields support $(\bar{\beta}(u)) \subseteq$ support $(\beta(u))$. $\qquad \square$

APPENDIX E: PROOF OF LEMMA 6 (USED IN THEOREM 5)

PROOF OF LEMMA 6. (Sparse Identifiability and Control of Empirical Error) The proof of claim (3.17) of this lemma follows identically the proof of claim (3.7) of Lemma 2, given in Appendix B, after replacing A_u with $\widetilde{A}_u$. Next we bound the emprical error

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

where the class of functions $\widetilde{\mathcal{F}}_{\widetilde{m}}$ is defined in Lemma 12. The result follows from the bound on $\epsilon_3(\widetilde{m})$ holding uniformly in $\widetilde{m} \leq n$ given in Lemma 13. $\qquad \square$

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

$$\mathcal{F}_{T,u} = \left\{ \alpha' x_i (1\{y_i \leq x_i'\beta\} - u) \ : \alpha \in \mathbb{S}(\beta), \text{support}(\beta) \subseteq T \right\}$$

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

$$\widetilde{\mathcal{F}}_{\widetilde{m}} = \left\{ \mathcal{F}_{T,u} : \ u \in \mathcal{U}, T \subset \{1, 2, \ldots, p\}, |T \setminus T_u| \leq \widetilde{m} \right\},$$

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

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

PROOF. The class $\mathcal{F}_{T,u}$ is a subset of $\mathcal{F}_T := \mathcal{G}_T + \mathcal{H}_T$ where $\mathcal{G}_T = \{\alpha' x_i \cdot 1\{y_i \leq x_i'\beta\} : \alpha \in \mathbb{S}(\beta), \text{support}(\beta) \subseteq T\}$ and $\mathcal{H}_T = \{-v \cdot \alpha' x_i : v \in \mathcal{U}, \alpha \in \mathbb{S}(\beta), \text{support}(\beta) \subseteq T\}$. The VC index of $\mathcal{G}_T$ and $\mathcal{H}_T$ is bounded by $c|T|$. Therefore the VC index of $\mathcal{F}_T$ is bounded by $2c|T| + 2 \leq V = 2c(\widetilde{m} + s) + 2$, for every $u \in \mathcal{U}$, which shows the first result.

To show the second result, we first note that the uniform covering numbers of $\mathcal{F}_T$ are bounded by $\sup_Q N(\epsilon \| F_T \|_{Q,2}, \mathcal{F}_T, L_2(Q)) \leq C(2V)(16e)^{2V+2}(1/\epsilon)^{2(2V-1)}$, where C is a universal constant (see [37] Theorem 2.6.7). We also note that $\widetilde{\mathcal{F}}_{\widetilde{m}}$ is a subset of $\widetilde{\mathcal{F}}_{\widetilde{m}}^o = \{\mathcal{F}_T : T \subset \{1, 2, \ldots, p\}, |T \setminus T_u| \leq \widetilde{m}, u \in \mathcal{U}\}$. Therefore, the bound stated in the lemma now follows by taking the product of the bound on the uniform covering numbers above with the total number of different function sets $\mathcal{F}_T$, indexed by models T, that generate $\widetilde{\mathcal{F}}_{\widetilde{m}}^o$, followed by some simplifications. To bound the number of function sets, first, note that for any fixed $u \in \mathcal{U}$, since $|T \setminus T_u| \leq \widetilde{m}$, we can pick at most $\max_{1 \leq k \leq \widetilde{m}} \binom{p}{k} \leq p^{\widetilde{m}}$ different models T; second, note that by varying across $u \in \mathcal{U}$, we can generate at most $\sum_{k=1}^{s} \binom{|\cup_{u \in \mathcal{U}} T_u|}{k} \leq \sum_{k=1}^{s} |\cup_{u \in \mathcal{U}} T_u|^k \leq 2|\cup_{u \in \mathcal{U}} T_u|^s$ different sets T_u since $s \leq |\cup_{u \in \mathcal{U}} T_u|$. The number of sets $\mathcal{F}_T$ is therefore bounded by $p^{\widetilde{m}} \cdot 2|\cup_{u \in \mathcal{U}} T_u|^s$. $\qquad\square$

LEMMA 13 (Controlling empirical error ϵ_3). *Suppose that D.1 holds and $|\cup_{u \in \mathcal{U}} T_u| \leq n$. There exists a universal constant A such that with probability at least $1 - \delta$,*

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

PROOF. Lemma 12 bounds the uniform covering number of $\widetilde{\mathcal{F}}_{\widetilde{m}}$. Using Lemma 17 with $m = \widetilde{m}$ and $\theta_m = p^2 \cdot n^{2s/\widetilde{m}}$, we conclude that uniformly in $\widetilde{m} \le n$

$$(\text{E.2}) \qquad \sup_{f \in \widetilde{\mathcal{F}}_{\widetilde{m}}} |\mathbb{G}_n(f)| \le A\delta^{-1/2}\sqrt{\widetilde{m}\log(n \vee \theta_m)} \cdot \max\left\{ \sup_{f \in \widetilde{\mathcal{F}}_{\widetilde{m}}} \mathrm{E}[f^2]^{1/2}, \sup_{f \in \widetilde{\mathcal{F}}_{\widetilde{m}}} \mathbb{E}_n[f^2]^{1/2} \right\}$$

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 $\phi(\widetilde{m} + s)$. $\qquad\square$

APPENDIX F: MAXIMAL INEQUALITIES FOR A COLLECTION OF EMPIRICAL PROCESSES

The main results here are Lemma 14 and Lemma 17, used in the proofs of Theorem 1 and Theorem 3 and 5, respectively. Lemma 17 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 17 by first stating Lemma 14, 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 16, 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 17.

LEMMA 14 (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 $g \in \mathcal{F} - \mathcal{F}$: $\mathbb{P}\{|\mathbb{G}(g)| > \eta\} \le 2\exp\left(-\frac{1}{2}\eta^2/D^2\|g\|_{P,2}^2\right)$ 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 $\mathcal{F}$:*

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

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* $c < 30$, $\rho(\mathcal{F}, P) := \sup_{f \in \mathcal{F}} \|f\|_{P,2}/\|F\|_{P,2}$,

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

The result of Lemma 14 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 $\tilde{v}, \tilde{w}$ created by the pairwise exchange of the components in v with components in w, we have that $x(\tilde{v}) \vee x(\tilde{w}) \ge [x(v) \vee x(w)]/k$. 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 [37]) to the case of a random threshold x that is sub-exchangeable.

LEMMA 15 (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* $x(Z)$, $\bar{p}_\tau := P(x(Z) \le q_\tau) \ge \tau$, *and* $p_\tau := P(x(Z) < q_\tau) \le \tau$. *Then*

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

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

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 16 (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*

$$\|F\|_{\mathbb{P}_n, 2} \int_0^{\rho(\mathcal{F}, \mathbb{P}_n)/4} \sqrt{\log n(\epsilon, \mathcal{F}, \mathbb{P}_n)} d\epsilon \le e_n(\mathcal{F}, \mathbb{P}_n) \text{ and } \sup_{f \in \mathcal{F}} var_{\mathbb{P}} f \le \frac{\tau}{2} (4kcKe_n(\mathcal{F}, \mathbb{P}_n))^2$$

for the same constant $c > 0$ as in Lemma 14, then

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

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

LEMMA 17 (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*

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

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

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

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

$$\sup_{f \in \mathcal{F}_m} |\mathbb{G}_n(f)| \leq 4\sqrt{2}cKe_n(\mathcal{F}_m, \mathbb{P}_n), \ \textit{for all } m \leq n,$$

with probability at least $1 - \delta$, provided that $n \vee \theta_0 \geq 3$; the constant c is the same as in Lemma 14.

PROOF OF LEMMA 14. The strategy of the proof is similar to the proof of Lemma 19.34 in [35], 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 q-th partition consists of sets of $L_2(P)$ radius at most $\|F\|_{P,2}2^{-q}$, where 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, e.g. [35], 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 $\cup_{q,i} f_{qi}$, 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

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

for constants η_q chosen below.

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

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

By Jensen $\sqrt{\log N_q} \leq a_q := \sum_{j=q_0}^q \sqrt{\log n_q}$, so that $\sum_{q=q_0+1}^{\infty} \eta_q \leq 8\sum_{q=q_0+1}^{\infty} K\|F\|_{P,2}2^{-q}a_q$. 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

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

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,

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

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

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

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 $\|\pi_{q_0}(f)\|_{P,2} \leq \|F\|_{P,2}$ and sub-Gaussianity, we conclude

$$\mathbb{P}\Big\{ \max_f |\mathbb{G}(\pi_{q_0}(f))| > \eta_{q_0} \Big\} \leq n_q 2\exp\left(-(K/D)^2 \log n_q\right) \leq 2\exp\left(-\{(K/D)^2 - 1\}\log n_q\right)$$

$$\leq \int_{q_0-1}^{q_0} 2\exp\left(-\{(K/D)^2 - 1\}\log n_q\right) dq = \int_{\rho(\mathcal{F},P)/4}^{\rho(\mathcal{F},P)/2} (x\ln 2)^{-1} 2n(x,\mathcal{F}, P)^{-\{(K/D)^2-1\}} dx.$$

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 $\sum_{q=q_0}^{\infty} \eta_q \leq cK\|F\|_{P,2} \int_0^{\rho(\mathcal{F},P)/4} \sqrt{\log n(x, \mathcal{F}, P)}dx$.

$\square$

PROOF OF LEMMA 15. The proof proceeds analogously to the proof of Lemma 2.3.7 (page 112) in [37] 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

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

By definition of x_0 we have $\inf_{f\in\mathcal{F}} P\left\{|\sum_{i=1}^n Y_i(f)| \leq \tfrac{x_0}{2}\right\} \geq 1 - \bar{p}_\tau/2$. 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 $\frac{\bar{p}_\tau}{2} \leq P_Y\left\{\|\sum_{i=1}^n (Y_i - Z_i)\|_{\mathcal{F}} > \tfrac{x_0 \vee x(Z) \vee x(Y)}{2}\right\}$. Integrating over Z we obtain

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

Let $\varepsilon_1, \ldots, \varepsilon_n$ be an independent sequence of Rademacher random variables. Given $\varepsilon_1, \ldots, \varepsilon_n$, set $(\tilde{Y}_i = Y_i, \tilde{Z}_i = Z_i)$ if $\varepsilon_i = 1$ and $(\tilde{Y}_i = Z_i, \tilde{Z}_i = Y_i)$ if $\varepsilon_i = -1$. That is, we create vectors

$\tilde{Y}$ and $\tilde{Z}$ by pairwise exchanging their components; by construction, conditional on each $\varepsilon_1, \ldots, \varepsilon_n$, $(\tilde{Y}, \tilde{Z})$ has the same distribution as (Y, Z). Therefore,

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

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

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

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

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

$\square$

PROOF OF LEMMA 16. 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, i.e., $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$, 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

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

Therefore, by the symmetrization Lemma 15 we obtain

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

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 $g \in \mathcal{F} - \mathcal{F}$, $\mathbb{P}_\varepsilon \{\mathbb{G}_n^o(g) > x\} \le 2 \exp(-x^2/[2\|g\|_{\mathbb{P}_n,2}^2])$. Hence by Lemma 14 with $D = 1$, we can bound

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

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

PROOF OF LEMMA 17. 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 16 uniformly in $m \le n$.

Second, recall $e_n(\mathcal{F}_m, \mathbb{P}_n) := C\sqrt{m\log(n \vee \theta_m \vee \omega)}\max\{\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P},2},\ \sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n,2}\}$ for $C = (1 + \sqrt{2v})/4$. 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 $m \leq n$:

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

which follows by $\int_0^\rho\sqrt{\log(\omega/\epsilon)}d\epsilon \leq (\int_0^\rho 1 d\epsilon)^{1/2}(\int_0^\rho \log(\omega/\epsilon)d\epsilon)^{1/2} \leq \rho\sqrt{2\log(n\vee\omega)}$, for $1/\sqrt{n} \leq \rho \leq 1$.

Third, for any $K \geq \sqrt{2/\delta} > 1$ we have $(K^2 - 1) \geq 1/\delta$, and let $\tau_m = \delta/(4m\log(n\vee\theta_0))$. Recall that $4\sqrt{2}cC > 4$ where $4 < c < 30$ is defined in Lemma 14. Note that for any $m \leq n$ and $f \in \mathcal{F}_m$, we have by Chebyshev inequality

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

By Lemma 16 with our choice of τ_m, $m \leq n$, $\omega > 1$, $v > 1$, and $\rho(\mathcal{F}_m, \mathbb{P}_n) \leq 1$,

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

where the last inequality follows by $n \vee \theta_0 \geq 3$ 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 $\tilde{Z}$ and $\tilde{Y}$ be created by exchanging any components in Z with corresponding compoents

in Y. Then

$$\sqrt{2}(\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(\tilde{Z}),2}\vee\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(\tilde{Y}),2})\geq(\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(\tilde{Z}),2}+\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(\tilde{Y}),2})^{1/2}$$
$$\geq(\sup_{f\in\mathcal{F}_m}\mathbb{E}_n[f(\tilde{Z}_i)^2]+\mathbb{E}_n[f(\tilde{Y}_i)^2])^{1/2}=(\sup_{f\in\mathcal{F}_m}\mathbb{E}_n[f(Z_i)^2]+\mathbb{E}_n[f(Y_i)^2])^{1/2}$$
$$\geq(\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(Z),2}\vee\sup_{f\in\mathcal{F}_m}\|f\|^2_{\mathbb{P}_n(Y),2})^{1/2}=\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(Z),2}\vee\sup_{f\in\mathcal{F}_m}\|f\|_{\mathbb{P}_n(Y),2}.$$

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 follows from $\mathbb{E}_n\left[F_m^2\right]=\mathbb{E}_n[\sup_{f\in\mathcal{F}_m}|f(Z_i)|^2]\leq\sup_{i\leq n}\sup_{f\in\mathcal{F}_m}|f(Z_i)|^2$, and from $\sup_{f\in\mathcal{F}_m}\mathbb{E}_n[|f(Z_i)|^2]\geq\sup_{f\in\mathcal{F}_m}\sup_{i\leq n}|f(Z_i)|^2/n$. $\qquad\square$

ACKNOWLEDGEMENTS

REFERENCES

[1] R. J. BARRO AND J.-W. LEE (1994). Data set for a panel of 139 countries, Discussion paper, NBER, http://www.nber.org/pub/barro.lee.

[2] R. J. BARRO AND X. SALA-I-MARTIN (1995). Economic Growth. McGraw-Hill, New York.

[3] A. BELLONI AND V. CHERNOZHUKOV (2008). Computational Complexity of MCMC-based Estimators in Large Samples, *The Annals of Statistics*, Volume 37, Number 4, 2011-2055.

[4] A. BELLONI AND V. CHERNOZHUKOV (2009). Supplemental Material for "ℓ_1-Penalized Quantile Regression in High-Dimensional Sparse Models", Duke and MIT Technical Report.

[5] A. BELLONI AND V. CHERNOZHUKOV (2008). Conditional Quantile Processes under Increasing Dimension, Duke and MIT Technical Report.

[6] D. BERTSIMAS AND J. TSITSIKLIS (1997). Introduction to Linear Optimization, Athena Scientific.

[7] P. J. BICKEL, Y. RITOV AND A. B. TSYBAKOV (2009). Simultaneous analysis of Lasso and Dantzig selector, The Annals of Statistics, forthcoming.

[8] M. BUCHINSKY (1994). Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression Econometrica, Vol. 62, No. 2 (Mar.), pp. 405–458.

[9] F. BUNEA, A. B. TSYBAKOV, AND M. H. WEGKAMP(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.

[10] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp (2007). Aggregation for Gaussian regression, The Annals of Statistics, Vol. 35, No. 4, 1674-1697.

[11] F. Bunea, A. Tsybakov, and M. H. Wegkamp (2007). Sparsity oracle inequalities for the Lasso, Electronic Journal of Statistics, Vol. 1, 169-194.

[12] E. Candes and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. Volume 35, Number 6, 2313-2351.

[13] V. Chernozhukov (2005). Extremal quantile regression. Ann. Statist. 33, no. 2, 806–839.

[14] R. Dudley (2000). Uniform Cental Limit Theorems, Cambridge Studies in advanced mathematics.

[15] J. Fan and J. Lv (2008). Sure Independence Screening for Ultra-High Dimensional Feature Space, Journal of the Royal Statistical Society Series B, 70, 849-911.

[16] C. Gutenbrunner and J. Jurečková (1992). Regression Rank Scores and Regression Quantiles The Annals of Statistics, Vol. 20, No. 1 (Mar.), pp. 305-330.

[17] X. He and Q.-M. Shao (2000). On Parameters of Increasing Dimenions, Journal of Multivariate Analysis, 73, 120-135.

[18] K. Knight (1998). Limiting distributions for L_1 regression estimators under general conditions, Annals of Statistics, 26, no. 2, 755–770.

[19] K. Knight and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. Ann. Statist. 28 1356-1378.

[20] R. Koenker (2005). Quantile regression, Econometric Society Monographs, Cambridge University Press.

[21] R. Koenker and G. Basset (1978). Regression Quantiles, Econometrica, Vol. 46, No. 1, January, 33–50.

[22] R. Koenker and J. Machado (1999). Goodness of fit and related inference process for quantile regression Journal of the American Statistical Association, 94, 1296–1310.

[23] V. Koltchinskii (2009). Sparsity in penalized empirical risk minimization, Ann. Inst. H. Poincar Probab. Statist. Volume 45, Number 1, 7-57.

[24] P.-S. Laplace (1818). *Théorie analytique des probabilités.* Éditions Jacques Gabay (1995), Paris.

[25] M. Ledoux and M. Talagrand (1991). Probability in Banach Spaces (Isoperimetry and processes). Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag.

[26] R. Levine and D. Renelt (1992). A Sensitivity Analysis of Cross-Country Growth Regressions, The American Economic Review, Vol. 82, No. 4, pp. 942-963.

[27] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer (2009). Taking Advantage of Sparsity in Multi-Task Learning, arXiv:0903.1468v1 [stat.ML].

[28] L. Lovász and S. Vempala(2007). The geometry of logconcave functions and sampling algorithms, Random Structures and Algorithms, Volume 30 Issue 3, pages 307 - 358.

[29] N. Meinshausen and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data, The Annals of Statistics, Vol. 37, No. 1, 246270.

[30] S. Portnoy (1991). Asymptotic behavior of regression quantiles in nonstationary, dependent cases. J. Multivariate Anal. 38, no. 1, 100–113.

[31] S. Portnoy and R. Koenker (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Statist. Sci. Volume 12, Number 4, 279-300.

[32] M. Rosenbaum and A. B. Tsybakov (2008). Sparse recovery under matrix uncertainty, arXiv:0812.2818v1 [math.ST].

[33] X. X. Sala-I-Martin(1997). I Just Ran Two Million Regressions, The American Economic Review, Vol. 87, No. 2, pp. 178-183.

[34] R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B, 58, 267-288.

[35] A. W. van der Vaart (1998). Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics.

[36] S. A. van de Geer (2008). High-dimensional generalized linear models and the Lasso, Annals of Statistics, Vol. 36, No. 2, 61–645.

[37] A. W. van der Vaart and J. A. Wellner (1996). Weak Convergence and Empirical Processes, Springer Series in Statistics.

[38] C.-H. Zhang and J. Huang (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann. Statist. Volume 36, Number 4, 1567-1594.

52

Alexandre Belloni
Duke University
Fuqua School of Business
1 Towerview Drive
Durham, NC 27708-0120
PO Box 90120
E-mail: abn5@duke.edu

Victor Chernozhukov
Massachusetts Institute of Technology
Department of Economics and Operations research Center
50 Memorial Drive
Room E52-262f
Cambridge, MA 02142
E-mail: vchern@mit.edu