UNIFORM POST SELECTION INFERENCE FOR LAD REGRESSION AND OTHER Z-ESTIMATION PROBLEMS

A. BELLONI, V. CHERNOZHUKOV, AND K. KATO

ABSTRACT. We develop uniformly valid confidence regions for regression coefficients in a high-dimensional sparse median regression model with homoscedastic errors. Our methods are based on a moment equation that is immunized against non-regular estimation of the nuisance part of the median regression function by using Neyman's orthogonalization. We establish that the resulting instrumental median regression estimator of a target regression coefficient is asymptotically normally distributed uniformly with respect to the underlying sparse model and is semi-parametrically efficient. We also generalize our method to a general non-smooth Z-estimation framework with the number of target parameters p_1 being possibly much larger than the sample size n. We extend Huber's results on asymptotic normality to this setting, demonstrating uniform asymptotic normality of the proposed estimators over p_1-dimensional rectangles, constructing simultaneous confidence bands on all of the p_1 target parameters, and establishing asymptotic validity of the bands uniformly over underlying approximately sparse models.

Keywords: Instrument; Post-selection inference; Sparsity; Neyman's Orthogonal Score test; Uniformly valid inference; Z-estimation.

Publication: Biometrika, 2014 doi:10.1093/biomet/asu056

1. INTRODUCTION

We consider independent and identically distributed data vectors $(y_i, x_i^{\mathrm{T}}, d_i)^{\mathrm{T}}$ that obey the regression model

$$y_i = d_i\alpha_0 + x_i^{\mathrm{T}}\beta_0 + \epsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where d_i is the main regressor and coefficient α_0 is the main parameter of interest. The vector x_i denotes other high-dimensional regressors or controls. The regression error ϵ_i is independent of d_i and x_i and has median zero, that is, $\mathrm{pr}(\epsilon_i \leq 0) = 1/2$. The distribution function of ϵ_i is denoted by F_ϵ and admits a density function f_ϵ such that $f_\epsilon(0) > 0$. The assumption motivates the use of the least absolute deviation or median regression, suitably adjusted for use in high-dimensional settings. The framework (1) is of interest in program evaluation, where d_i represents the treatment or policy variable known a priori and whose impact we would like

Date: First version: May 2012, this version May, 2014. We would like to thank the participants of Luminy conference on Nonparametric and high-dimensional statistics (December 2012), Oberwolfach workshop on Frontiers in Quantile Regression (November 2012), 8th World Congress in Probability and Statistics (August 2012), and seminar at the University of Michigan (October 2012). This paper was first presented in 8th World Congress in Probability and Statistics in August 2012. We would like to thank the editor, an associate editor, and anonymous referees for their careful review. We are also grateful to Sara van de Geer, Xuming He, Richard Nickl, Roger Koenker, Vladimir Koltchinskii, Enno Mammen, Steve Portnoy, Philippe Rigollet, Richard Samworth, and Bin Yu for useful comments and discussions. Research support from the National Science Foundation and the Japan Society for the Promotion of Science is gratefully acknowledged.

to infer [27, 21, 15]. We shall also discuss a generalization to the case where there are many parameters of interest, including the case where the identity of a regressor of interest is unknown a priori.

The dimension p of controls x_i may be much larger than n, which creates a challenge for inference on α_0. Although the unknown nuisance parameter β_0 lies in this large space, the key assumption that will make estimation possible is its sparsity, namely $T = \mathrm{supp}(\beta_0)$ has $s < n$ elements, where the notation $\mathrm{supp}(\delta) = \{j \in \{1, \ldots, p\} : \delta_j \neq 0\}$ denotes the support of a vector $\delta \in \mathbb{R}^p$. Here s can depend on n, as we shall use array asymptotics. Sparsity motivates the use of regularization or model selection methods.

A non-robust approach to inference in this setting would be first to perform model selection via the ℓ_1-penalized median regression estimator

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} E_n(|y_i - d_i\alpha - x_i^{\mathrm{T}}\beta|) + \frac{\lambda}{n}\|\Psi(\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}\|_1, \tag{2}$$

where λ is a penalty parameter and $\Psi^2 = \mathrm{diag}\{E_n(d_i^2), E_n(x_{i1}^2), \ldots, E_n(x_{ip}^2)\}$ is a diagonal matrix with normalization weights, where the notation $E_n(\cdot)$ denotes the average $n^{-1}\sum_{i=1}^n$ over the index $i = 1, \ldots, n$. Then one would use the post-model selection estimator

$$(\widetilde{\alpha}, \widetilde{\beta}) \in \arg\min_{\alpha,\beta} \left\{ E_n(|y_i - d_i\alpha - x_i^{\mathrm{T}}\beta|) : \ \beta_j = 0, \ j \notin \mathrm{supp}(\widehat{\beta}) \right\}, \tag{3}$$

to perform inference for α_0.

This approach is justified if (2) achieves perfect model selection with probability approaching unity, so that the estimator (3) has the oracle property. However conditions for perfect selection are very restrictive in this model, and, in particular, require strong separation of non-zero coefficients away from zero. If these conditions do not hold, the estimator $\widetilde{\alpha}$ does not converge to α_0 at the $n^{-1/2}$ rate, uniformly with respect to the underlying model, and so the usual inference breaks down [19]. We shall demonstrate the breakdown of such naive inference in Monte Carlo experiments where non-zero coefficients in β_0 are not significantly separated from zero.

The breakdown of standard inference does not mean that the aforementioned procedures are not suitable for prediction. Indeed, the estimators (2) and (3) attain essentially optimal rates $\{(s \log p)/n\}^{1/2}$ of convergence for estimating the entire median regression function [3, 33]. This property means that while these procedures will not deliver perfect model recovery, they will only make moderate selection mistakes, that is, they omit controls only if coefficients are local to zero.

In order to provide uniformly valid inference, we propose a method whose performance does not require perfect model selection, allowing potential moderate model selection mistakes. The latter feature is critical in achieving uniformity over a large class of data generating processes, similarly to the results for instrumental regression and mean regression studied in [34] and [2, 4, 5]. This allows us to overcome the impact of moderate model selection mistakes on inference, avoiding in part the criticisms in [19], who prove that the oracle property achieved by the naive estimators implies the failure of uniform validity of inference and their semiparametric inefficiency [20].

In order to achieve robustness with respect to moderate selection mistakes, we shall construct an orthogonal moment equation that identifies the target parameter. The following auxiliary equation,

$$d_i = x_i^{\mathrm{T}}\theta_0 + v_i, \quad E(v_i \mid x_i) = 0 \quad (i = 1, \ldots, n), \tag{4}$$

which describes the dependence of the regressor of interest d_i on the other controls x_i, plays a key role. We shall assume the sparsity of θ_0, that is, $T_d = \mathrm{supp}(\theta_0)$ has at most $s < n$ elements, and estimate the relation (4) via lasso or post-lasso least squares methods described below.

We shall use v_i as an instrument in the following moment equation for α_0:

$$E\{\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta_0)v_i\} = 0 \quad (i = 1, \ldots, n), \tag{5}$$

where $\varphi(t) = 1/2 - 1\{t \leq 0\}$. We shall use the empirical analog of (5) to form an instrumental median regression estimator of α_0, using a plug-in estimator for $x_i^{\mathrm{T}}\beta_0$. The moment equation (5) has the orthogonality property

$$\frac{\partial}{\partial\beta}E\{\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta)v_i\}\Big|_{\beta=\beta_0} = 0 \quad (i = 1, \ldots, n), \tag{6}$$

so the estimator of α_0 will be unaffected by estimation of $x_i^{\mathrm{T}}\beta_0$ even if β_0 is estimated at a slower rate than $n^{-1/2}$, that is, the rate of $o(n^{-1/4})$ would suffice. This slow rate of estimation of the nuisance function permits the use of non-regular estimators of β_0, such as post-selection or regularized estimators that are not $n^{-1/2}$ consistent uniformly over the underlying model. The orthogonalization ideas can be traced back to [22] and also play an important role in doubly robust estimation [26].

Our estimation procedure has three steps: (i) estimation of the confounding function $x_i^{\mathrm{T}}\beta_0$ in (1); (ii) estimation of the instruments v_i in (4); and (iii) estimation of the target parameter α_0 via empirical analog of (5). Each step is computationally tractable, involving solutions of convex problems and a one-dimensional search.

Step (i) estimates for the nuisance function $x_i^{\mathrm{T}}\beta_0$ via either the ℓ_1-penalized median regression estimator (2) or the associated post-model selection estimator (3).

Step (ii) provides estimates $\widehat{v}_i$ of v_i in (4) as $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widehat{\theta}$ or $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}$ $(i = 1, \ldots, n)$. The first is based on the heteroscedastic lasso estimator $\widehat{\theta}$, a version of the lasso of [30], designed to address non-Gaussian and heteroscedastic errors [2],

$$\widehat{\theta} \in \arg\min_{\theta} E_n\{(d_i - x_i^{\mathrm{T}}\theta)^2\} + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1, \tag{7}$$

where λ and $\widehat{\Gamma}$ are the penalty level and data-driven penalty loadings defined in the Supplementary Material. The second is based on the associated post-model selection estimator and $\widetilde{\theta}$, called the post-lasso estimator:

$$\widetilde{\theta} \in \arg\min_{\theta}\left[E_n\{(d_i - x_i^{\mathrm{T}}\theta)^2\} : \ \theta_j = 0, \ j \notin \mathrm{supp}(\widehat{\theta})\right]. \tag{8}$$

Step (iii) constructs an estimator $\check{\alpha}$ of the coefficient α_0 via an instrumental median regression [11], using $(\widehat{v}_i)_{i=1}^n$ as instruments, defined by

$$\check{\alpha} \in \arg\min_{\alpha\in\widehat{\mathcal{A}}} L_n(\alpha), \quad L_n(\alpha) = \frac{4|E_n\{\varphi(y_i - x_i^{\mathrm{T}}\widehat{\beta} - d_i\alpha)\widehat{v}_i\}|^2}{E_n(\widehat{v}_i^2)}, \tag{9}$$

where $\widehat{\mathcal{A}}$ is a possibly stochastic parameter space for α_0. We suggest $\widehat{\mathcal{A}} = [\widehat{\alpha} - 10/b, \widehat{\alpha} + 10/b]$ with $b = \{E_n(d_i^2)\}^{1/2}\log n$, though we allow for other choices.

Our main result establishes that under homoscedasticity, provided that $(s^3\log^3 p)/n \to 0$ and other regularity conditions hold, despite possible model selection mistakes in Steps (i) and (ii),

the estimator $\breve{\alpha}$ obeys

$$\sigma_n^{-1} n^{1/2} (\breve{\alpha} - \alpha_0) \to N(0,1) \tag{10}$$

in distribution, where $\sigma_n^2 = 1/\{4 f_\epsilon^2 E(v_i^2)\}$ with $f_\epsilon = f_\epsilon(0)$ is the semi-parametric efficiency bound for regular estimators of α_0. In the low-dimensional case, if $p^3 = o(n)$, the asymptotic behavior of our estimator coincides with that of the standard median regression without selection or penalization, as derived in [13], which is also semi-parametrically efficient in this case. However, the behaviors of our estimator and the standard median regression differ dramatically, otherwise, with the standard estimator even failing to be consistent when $p > n$. Of course, this improvement in the performance comes at the cost of assuming sparsity.

An alternative, more robust expression for σ_n^2 is given by

$$\sigma_n^2 = J^{-1} \Omega J^{-1}, \quad \Omega = E(v_i^2)/4, \quad J = E(f_\epsilon d_i v_i). \tag{11}$$

We estimate Ω by the plug-in method and J by Powell's ([25]) method. Furthermore, we show that the Neyman-type projected score statistic $nL_n(\alpha)$ can be used for testing the null hypothesis $\alpha = \alpha_0$, and converges in distribution to a χ_1^2 variable under the null hypothesis, that is,

$$nL_n(\alpha_0) \to \chi_1^2 \tag{12}$$

in distribution. This allows us to construct a confidence region with asymptotic coverage $1 - \xi$ based on inverting the score statistic $nL_n(\alpha)$:

$$\widehat{A}_\xi = \{\alpha \in \widehat{\mathcal{A}} : nL_n(\alpha) \leq q_{1-\xi}\}, \quad \mathrm{pr}(\alpha_0 \in \widehat{A}_\xi) \to 1 - \xi, \tag{13}$$

where $q_{1-\xi}$ is the $(1-\xi)$-quantile of the χ_1^2-distribution.

The robustness with respect to moderate model selection mistakes, which is due to (6), allows (10) and (12) to hold uniformly over a large class of data generating processes. Throughout the paper, we use array asymptotics, asymptotics where the model changes with n, to better capture finite-sample phenomena such as small coefficients that are local to zero. This ensures the robustness of conclusions with respect to perturbations of the data-generating process along various model sequences. This robustness, in turn, translates into uniform validity of confidence regions over many data-generating processes.

The second set of main results addresses a more general setting by allowing p_1-dimensional target parameters defined via Huber's Z-problems to be of interest, with dimension p_1 potentially much larger than the sample size n, and also allowing for approximately sparse models instead of exactly sparse models. This framework covers a wide variety of semi-parametric models, including those with smooth and non-smooth score functions. We provide sufficient conditions to derive a uniform Bahadur representation, and establish uniform asymptotic normality, using central limit theorems and bootstrap results of [9], for the entire p_1-dimensional vector. The latter result holds uniformly over high-dimensional rectangles of dimension $p_1 \gg n$ and over an underlying approximately sparse model, thereby extending previous results from the setting with $p_1 \ll n$ [14, 23, 24, 13] to that with $p_1 \gg n$.

In what follows, the ℓ_2 and ℓ_1 norms are denoted by $\|\cdot\|$ and $\|\cdot\|_1$, respectively, and the ℓ_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. We use the notation $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. Denote by $\Phi(\cdot)$ the distribution function of the standard normal distribution. We assume that the quantities such as p, s, and hence $y_i, x_i, \beta_0, \theta_0, T$ and T_d are all dependent on the sample size n, and allow for the case where $p = p_n \to \infty$ and $s = s_n \to \infty$ as $n \to \infty$. We shall omit the dependence of these quantities on n when it does not cause

confusion. For a class of measurable functions $\mathcal{F}$ on a measurable space, let $\mathrm{cn}(\epsilon, \mathcal{F}, \|\cdot\|_{Q,2})$ denote its ϵ-covering number with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$, where Q is a finitely discrete measure on the space, and let $\mathrm{ent}(\varepsilon, \mathcal{F}) = \log \sup_Q \mathrm{cn}(\varepsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2})$ denote the uniform entropy number where $F = \sup_{f\in\mathcal{F}} |f|$.

2. The Methods, Conditions, and Results

2.1. **The methods.** Each of the steps outlined in Section 1 could be implemented by several estimators. Two possible implementations are the following.

Algorithm 1. *The algorithm is based on post-model selection estimators.*
Step *(i). Run post-ℓ_1-penalized median regression (3) of y_i on d_i and x_i; keep fitted value $x_i^{\mathrm{T}}\widetilde{\beta}$.*
Step *(ii). Run the post-lasso estimator (8) of d_i on x_i; keep the residual $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}$.*
Step *(iii). Run instrumental median regression (9) of $y_i - x_i^{\mathrm{T}}\widetilde{\beta}$ on d_i using $\widehat{v}_i$ as the instrument. Report $\check{\alpha}$ and perform inference based upon (10) or (13).*

Algorithm 2. *The algorithm is based on regularized estimators.*
Step *(i). Run ℓ_1-penalized median regression (3) of y_i on d_i and x_i; keep fitted value $x_i^{\mathrm{T}}\widetilde{\beta}$.*
Step *(ii). Run the lasso estimator (7) of d_i on x_i; keep the residual $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}$.*
Step *(iii). Run instrumental median regression (9) of $y_i - x_i^{\mathrm{T}}\widetilde{\beta}$ on d_i using $\widehat{v}_i$ as the instrument. Report $\check{\alpha}$ and perform inference based upon (10) or (13).*

In order to perform ℓ_1-penalized median regression and lasso, one has to choose the penalty levels suitably. We record our penalty choices in the Supplementary Material. Algorithm 1 relies on the post-selection estimators that refit the non-zero coefficients without the penalty term to reduce the bias, while Algorithm 2 relies on the penalized estimators. In Step (ii), instead of the lasso or the post-lasso estimators, Dantzig selector [8] and Gauss-Dantzig estimators could be used. Step (iii) of both algorithms relies on instrumental median regression (9).

Comment 2.1. Alternatively, in this step, we can use a one-step estimator $\check{\alpha}$ defined by

$$\check{\alpha} = \widehat{\alpha} + [E_n\{f_\epsilon(0)\widehat{v}_i^2\}]^{-1} E_n\{\varphi(y_i - d_i\widehat{\alpha} - x_i^{\mathrm{T}}\widehat{\beta})\widehat{v}_i\}, \tag{14}$$

where $\widehat{\alpha}$ is the ℓ_1-penalized median regression estimator (2). Another possibility is to use the post-double selection median regression estimation, which is simply the median regression of y_i on d_i and the union of controls selected in both Steps (i) and (ii), as $\check{\alpha}$. The Supplemental Material shows that these alternative estimators also solve (9) approximately.

2.2. **Regularity conditions.** We state regularity conditions sufficient for validity of the main estimation and inference results. The behavior of sparse eigenvalues of the population Gram matrix $E(\widetilde{x}_i\widetilde{x}_i^{\mathrm{T}})$ with $\widetilde{x}_i = (d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$ plays an important role in the analysis of ℓ_1-penalized median regression and lasso. Define the minimal and maximal m-sparse eigenvalues of the population Gram matrix as

$$\bar{\phi}_{\min}(m) = \min_{1\le\|\delta\|_0\le m} \frac{\delta^{\mathrm{T}} E(\widetilde{x}_i\widetilde{x}_i^{\mathrm{T}})\delta}{\|\delta\|^2}, \quad \bar{\phi}_{\max}(m) = \max_{1\le\|\delta\|_0\le m} \frac{\delta^{\mathrm{T}} E(\widetilde{x}_i\widetilde{x}_i^{\mathrm{T}})\delta}{\|\delta\|^2}, \tag{15}$$

where $m = 1, \ldots, p$. Assuming that $\bar{\phi}_{\min}(m) > 0$ requires that all population Gram submatrices formed by any m components of $\widetilde{x}_i$ are positive definite.

The main condition, Condition 1, imposes sparsity of the vectors β_0 and θ_0 as well as other more technical assumptions. Below let c_1 and C_1 be given positive constants, and let $\ell_n \uparrow \infty, \delta_n \downarrow 0$, and $\Delta_n \downarrow 0$ be given sequences of positive constants.

Condition 1. *Suppose that (i)* $\{(y_i, d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ *is a sequence of independent and identically distributed random vectors generated according to models (1) and (4), where* ϵ_i *has distribution distribution function* F_ϵ *such that* $F_\epsilon(0) = 1/2$ *and is independent of the random vector* $(d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$*; (ii)* $E(v_i^2 \mid x) \geq c_1$ *and* $E(|v_i|^3 \mid x_i) \leq C_1$ *almost surely; moreover,* $E(d_i^4) + E(v_i^4) + \max_{j=1,\ldots,p} E(x_{ij}^2 d_i^2) + E(|x_{ij} v_i|^3) \leq C_1$*; (iii) there exists* $s = s_n \geq 1$ *such that* $\|\beta_0\|_0 \leq s$ *and* $\|\theta_0\|_0 \leq s$*; (iv) the error distribution* F_ϵ *is absolutely continuous with continuously differentiable density* $f_\epsilon(\cdot)$ *such that* $f_\epsilon(0) \geq c_1$ *and* $f_\epsilon(t) \vee |f_\epsilon'(t)| \leq C_1$ *for all* $t \in \mathbb{R}$*; (v) there exist constants* K_n *and* M_n *such that* $K_n \geq \max_{j=1,\ldots,p} |x_{ij}|$ *and* $M_n \geq 1 \vee |x_i^{\mathrm{T}} \theta_0|$ *almost surely, and they obey the growth condition* $\{K_n^4 + (K_n^2 \vee M_n^4) s^2 + M_n^2 s^3\} \log^3(p \vee n) \leq n \delta_n$*; (vi)* $c_1 \leq \bar{\phi}_{\min}(\ell_n s) \leq \bar{\phi}_{\max}(\ell_n s) \leq C_1$.

Condition 1 (i) imposes the setting discussed in the previous section with the zero conditional median of the error distribution. Condition 1 (ii) imposes moment conditions on the structural errors and regressors to ensure good model selection performance of lasso applied to equation (4). Condition 1 (iii) imposes sparsity of the high-dimensional vectors β_0 and θ_0. Condition 1 (iv) is a set of standard assumptions in median regression [16] and in instrumental quantile regression. Condition 1 (v) restricts the sparsity index, namely $s^3 \log^3(p \vee n) = o(n)$ is required; this is analogous to the restriction $p^3 (\log p)^2 = o(n)$ made in [13] in the low-dimensional setting. The uniformly bounded regressors condition can be relaxed with minor modifications provided the bound holds with probability approaching unity. Most importantly, no assumptions on the separation from zero of the non-zero coefficients of θ_0 and β_0 are made. Condition 1 (vi) is quite plausible for many designs of interest. Conditions 1 (iv) and (v) imply the equivalence between the norms induced by the empirical and population Gram matrices over s-sparse vectors by [29].

2.3. **Results.** The following result is derived as an application of a more general Theorem 2 given in Section 3; the proof is given in the Supplementary Material.

Theorem 1. *Let* $\check{\alpha}$ *and* $L_n(\alpha_0)$ *be the estimator and statistic obtained by applying either Algorithm 1 or 2. Suppose that Condition 1 is satisfied for all* $n \geq 1$*. Moreover, suppose that with probability at least* $1 - \Delta_n$, $\|\widehat{\beta}\|_0 \leq C_1 s$*. Then, as* $n \to \infty$, $\sigma_n^{-1} n^{1/2} (\check{\alpha} - \alpha_0) \to N(0,1)$ *and* $n L_n(\alpha_0) \to \chi_1^2$ *in distribution, where* $\sigma_n^2 = 1/\{4 f_\epsilon^2 E(v_i^2)\}$.

Theorem 1 shows that Algorithms 1 and 2 produce estimators $\check{\alpha}$ that perform equally well, to the first order, with asymptotic variance equal to the semi-parametric efficiency bound; see the Supplemental Material for further discussion. Both algorithms rely on sparsity of $\widehat{\beta}$ and $\widehat{\theta}$. Sparsity of the latter follows immediately under sharp penalty choices for optimal rates. The sparsity for the former potentially requires a higher penalty level, as shown in [3]; alternatively, sparsity for the estimator in Step 1 can also be achieved by truncating the smallest components of $\widehat{\beta}$. The Supplemental Material shows that suitable truncation leads to the required sparsity while preserving the rate of convergence.

An important consequence of these results is the following corollary. Here $\mathcal{P}_n$ denotes a collection of distributions for $\{(y_i, d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ and for $P_n \in \mathcal{P}_n$ the notation pr_{P_n} means that under pr_{P_n}, $\{(y_i, d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ is distributed according to the law determined by P_n.

Corollary 1. *Let $\breve{\alpha}$ be the estimator of α_0 constructed according to either Algorithm 1 or 2, and for every $n \geq 1$, let $\mathcal{P}_n$ be the collection of all distributions of $\{(y_i, d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ for which Condition 1 holds and $\|\widehat{\beta}\|_0 \leq C_1 s$ with probability at least $1 - \Delta_n$. Then for $\widehat{A}_\xi$ defined in (13),*

$$\sup_{P_n \in \mathcal{P}_n} \left| \mathrm{pr}_{P_n} \left\{ \alpha_0 \in [\breve{\alpha} \pm \sigma_n n^{-1/2} \Phi^{-1}(1 - \xi/2)] \right\} - (1 - \xi) \right| \to 0,$$

$$\sup_{P_n \in \mathcal{P}_n} \left| \mathrm{pr}_{P_n}(\alpha_0 \in \widehat{A}_\xi) - (1 - \xi) \right| \to 0, \quad n \to \infty.$$

Corollary 1 establishes the second main result of the paper. It highlights the uniform validity of the results, which hold despite the possible imperfect model selection in Steps (i) and (ii). Condition 1 explicitly characterizes regions of data-generating processes for which the uniformity result holds. Simulations presented below provide additional evidence that these regions are substantial. Here we rely on exactly sparse models, but these results extend to approximately sparse model in what follows.

Both of the proposed algorithms exploit the homoscedasticity of the model (1) with respect to the error term ϵ_i. The generalization to the heteroscedastic case can be achieved but we need to consider the density-weighted version of the auxiliary equation (4) in order to achieve the semiparametric efficiency bound. The analysis of the impact of estimation of weights is delicate and is developed in our working paper "Robust Inference in High-Dimensional Approximate Sparse Quantile Regression Models" (arXiv:1312.7186).

2.4. **Generalization to many target coefficients.** We consider the generalization to the previous model:

$$y = \sum_{j=1}^{p_1} d_j \alpha_j + g(u) + \epsilon, \quad \epsilon \sim F_\epsilon, \quad F_\epsilon(0) = 1/2,$$

where d, u are regressors, and ϵ is the noise with distribution function F_ϵ that is independent of regressors and has median zero, that is, $F_\epsilon(0) = 1/2$. The coefficients $\alpha_1, \ldots, \alpha_{p_1}$ are now the high-dimensional parameter of interest.

We can rewrite this model as p_1 models of the previous form:

$$y = \alpha_j d_j + g_j(z_j) + \epsilon, \quad d_j = m_j(z_j) + v_j, \quad E(v_j \mid z_j) = 0 \quad (j = 1, \ldots, p_1),$$

where α_j is the target coefficient,

$$g_j(z_j) = \sum_{k \neq j}^{p_1} d_k \alpha_k + g(u), \quad m_j(z_j) = E(d_j \mid z_j),$$

and where $z_j = (d_1, \ldots, d_{j-1}, d_{j+1}, \ldots, d_{p_1}, u^{\mathrm{T}})^{\mathrm{T}}$. We would like to estimate and perform inference on each of the p_1 coefficients $\alpha_1, \ldots, \alpha_{p_1}$ simultaneously.

Moreover, we would like to allow regression functions $h_j = (g_j, m_j)^{\mathrm{T}}$ to be of infinite dimension, that is, they could be written only as infinite linear combinations of some dictionary with respect to z_j. However, we assume that there are sparse estimators $\widehat{h}_j = (\widehat{g}_j, \widehat{m}_j)^{\mathrm{T}}$ that can estimate $h_j = (g_j, m_j)^{\mathrm{T}}$ at sufficiently fast $o(n^{-1/4})$ rates in the mean square error sense, as stated precisely in Section 3. Examples of functions h_j that permit such estimation by sparse methods include the standard Sobolev spaces as well as more general rearranged Sobolev spaces [7, 6]. Here sparsity of estimators $\widehat{g}_j$ and $\widehat{m}_j$ means that they are formed by $O_P(s)$-sparse linear

combinations chosen from p technical regressors generated from z_j, with coefficients estimated from the data. This framework is general; in particular it contains as a special case the traditional linear sieve/series framework for estimation of h_j, which uses a small number $s = o(n)$ of predetermined series functions as a dictionary.

Given suitable estimators for $h_j = (g_j, m_j)^{\mathrm{T}}$, we can then identify and estimate each of the target parameters $(\alpha_j)_{j=1}^{p_1}$ via the empirical version of the moment equations

$$E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = 0 \quad (j = 1, \ldots, p_1),$$

where $\psi_j(w, \alpha, t) = \varphi(y - d_j\alpha - t_1)(d_j - t_2)$ and $w = (y, d_1, \ldots, d_{p_1}, u^{\mathrm{T}})^{\mathrm{T}}$. These equations have the orthogonality property:

$$[\partial E\{\psi_j(w, \alpha_j, t) \mid z_j\}/\partial t]\big|_{t=h_j(z_j)} = 0 \quad (j = 1, \ldots, p_1).$$

The resulting estimation problem is subsumed as a special case in the next section.

3. Inference on many target parameters in Z-problems

In this section we generalize the previous example to a more general setting, where p_1 target parameters defined via Huber's Z-problems are of interest, with dimension p_1 potentially much larger than the sample size. This framework covers median regression, its generalization discussed above, and many other semi-parametric models.

The interest lies in $p_1 = p_{1n}$ real-valued target parameters $\alpha_1, \ldots, \alpha_{p_1}$. We assume that each $\alpha_j \in \mathcal{A}_j$, where each $\mathcal{A}_j$ is a non-stochastic bounded closed interval. The true parameter α_j is identified as a unique solution of the moment condition:

$$E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = 0. \tag{16}$$

Here w is a random vector taking values in $\mathcal{W}$, a Borel subset of a Euclidean space, which contains vectors z_j $(j = 1, \ldots, p_1)$ as subvectors, and each z_j takes values in $\mathcal{Z}_j$; here z_j and $z_{j'}$ with $j \neq j'$ may overlap. The vector-valued function $z \mapsto h_j(z) = \{h_{jm}(z)\}_{m=1}^M$ is a measurable map from $\mathcal{Z}_j$ to $\mathbb{R}^M$, where M is fixed, and the function $(w, \alpha, t) \mapsto \psi_j(w, \alpha, t)$ is a measurable map from an open neighborhood of $\mathcal{W} \times \mathcal{A}_j \times \mathbb{R}^M$ to $\mathbb{R}$. The former map is a possibly infinite-dimensional nuisance parameter.

Suppose that the nuisance function $h_j = (h_{jm})_{m=1}^M$ admits a sparse estimator $\widehat{h}_j = (\widehat{h}_{jm})_{m=1}^M$ of the form

$$\widehat{h}_{jm}(\cdot) = \sum_{k=1}^{p} f_{jmk}(\cdot)\widehat{\theta}_{jmk}, \quad \|(\widehat{\theta}_{jmk})_{k=1}^p\|_0 \leq s \quad (m = 1, \ldots, M),$$

where $p = p_n$ may be much larger than n while $s = s_n$, the sparsity level of $\widehat{h}_j$, is small compared to n, and $f_{jmk} : \mathcal{Z}_j \to \mathbb{R}$ are given approximating functions.

The estimator $\widehat{\alpha}_j$ of α_j is then constructed as a Z-estimator, which solves the sample analogue of the equation (16):

$$|E_n[\psi_j\{w, \widehat{\alpha}_j, \widehat{h}_j(z_j)\}]| \leq \inf_{\alpha \in \widehat{\mathcal{A}}_j} |E_n[\psi\{w, \alpha, \widehat{h}_j(z_j)\}]| + \epsilon_n, \tag{17}$$

where $\epsilon_n = o(n^{-1/2}b_n^{-1})$ is the numerical tolerance parameter and $b_n = \{\log(ep_1)\}^{1/2}$; $\widehat{\mathcal{A}}_j$ is a possibly stochastic interval contained in $\mathcal{A}_j$ with high probability. Typically, $\widehat{\mathcal{A}}_j = \mathcal{A}_j$ or can be constructed by using a preliminary estimator of α_j.

In order to achieve robust inference results, we shall need to rely on the condition of orthogonality, or immunity, of the scores with respect to small perturbations in the value of the nuisance parameters, which we can express in the following condition:

$$\partial_t E\{\psi_j(w, \alpha_j, t) \mid z_j\}|_{t=h_j(z_j)} = 0, \tag{18}$$

where we use the symbol ∂_t to abbreviate $\partial/\partial t$. It is important to construct the scores ψ_j to have property (18) or its generalization given in Remark 3.1 below. Generally, we can construct the scores ψ_j that obey such properties by projecting some initial non-orthogonal scores onto the orthogonal complement of the tangent space for the nuisance parameter [**?**,]]vdV-W,vdV,kosorok:book. Sometimes the resulting construction generates additional nuisance parameters, for example, the auxiliary regression function in the case of the median regression problem in Section 2.

In Conditions 2 and 3 below, ς, n_0, c_1, and C_1 are given positive constants; M is a fixed positive integer; $\delta_n \downarrow 0$ and $\rho_n \downarrow 0$ are given sequences of constants. Let $a_n = \max(p_1, p, n, \mathrm{e})$ and $b_n = \{\log(ep_1)\}^{1/2}$.

Condition 2. *For every $n \geq 1$, we observe independent and identically distributed copies $(w_i)_{i=1}^n$ of the random vector w, whose law is determined by the probability measure $P \in \mathcal{P}_n$. Uniformly in $n \geq n_0, P \in \mathcal{P}_n$, and $j = 1, \ldots, p_1$, the following conditions are satisfied: (i) the true parameter α_j obeys (16); $\widehat{\mathcal{A}}_j$ is a possibly stochastic interval such that with probability $1 - \delta_n$, $[\alpha_j \pm c_1 n^{-1/2} \log^2 a_n] \subset \widehat{\mathcal{A}}_j \subset \mathcal{A}_j$; (ii) for P-almost every z_j, the map $(\alpha, t) \mapsto E\{\psi_j(w, \alpha, t) \mid z_j\}$ is twice continuously differentiable, and for every $\nu \in \{\alpha, t_1, \ldots, t_M\}$, $E(\sup_{\alpha_j \in \mathcal{A}_j} |\partial_\nu E[\psi_j\{w, \alpha, h_j(z_j)\} \mid z_j]|^2) \leq C_1$; moreover, there exist constants $L_{1n} \geq 1, L_{2n} \geq 1$, and a cube $\mathcal{T}_j(z_j) = \times_{m=1}^M \mathcal{T}_{jm}(z_j)$ in $\mathbb{R}^M$ with center $h_j(z_j)$ such that for every $\nu, \nu' \in \{\alpha, t_1, \ldots, t_M\}$, $\sup_{(\alpha,t) \in \mathcal{A}_j \times \mathcal{T}_j(z_j)} |\partial_\nu \partial_{\nu'} E\{\psi_j(w, \alpha, t) \mid z_j\}| \leq L_{1n}$, and for every $\alpha, \alpha' \in \mathcal{A}_j, t, t' \in \mathcal{T}_j(z_j)$, $E[\{\psi_j(w, \alpha, t) - \psi_j(w, \alpha', t')\}^2 \mid z_j] \leq L_{2n}(|\alpha - \alpha'|^\varsigma + \|t - t'\|^\varsigma)$; (iii) the orthogonality condition (18) or its generalization stated in (20) below holds; (iv) the following global and local identifiability conditions hold: $2|E[\psi_j\{w, \alpha, h_j(z_j)\}]| \geq |\Gamma_j(\alpha - \alpha_j)| \wedge c_1$ for all $\alpha \in \mathcal{A}_j$, where $\Gamma_j = \partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\}]$, and $|\Gamma_j| \geq c_1$; and (v) the second moments of scores are bounded away from zero: $E[\psi_j^2\{w, \alpha_j, h_j(z_j)\}] \geq c_1$.*

Condition 2 states rather mild assumptions for Z-estimation problems, in particular, allowing for non-smooth scores ψ_j such as those arising in median regression. They are analogous to assumptions imposed in the setting with $p = o(n)$, for example, in [13]. The following condition uses a notion of pointwise measurable classes of functions [32].

Condition 3. *Uniformly in $n \geq n_0, P \in \mathcal{P}_n$, and $j = 1, \ldots, p_1$, the following conditions are satisfied: (i) the nuisance function $h_j = (h_{jm})_{m=1}^M$ has an estimator $\widehat{h}_j = (\widehat{h}_{jm})_{m=1}^M$ with good sparsity and rate properties, namely, with probability $1 - \delta_n$, $\widehat{h}_j \in \mathcal{H}_j$, where $\mathcal{H}_j = \times_{m=1}^M \mathcal{H}_{jm}$ and each $\mathcal{H}_{jm}$ is the class of functions $\widetilde{h}_{jm} : \mathcal{Z}_j \to \mathbb{R}$ of the form $\widetilde{h}_{jm}(\cdot) = \sum_{k=1}^p f_{jmk}(\cdot)\theta_{mk}$ such that $\|(\theta_{mk})_{k=1}^p\|_0 \leq s$, $\widetilde{h}_{jm}(z) \in \mathcal{T}_{jm}(z)$ for all $z \in \mathcal{Z}_j$, and $E[\{\widetilde{h}_{jm}(z_j) - h_{jm}(z_j)\}^2] \leq C_1 s(\log a_n)/n$, where $s = s_n \geq 1$ is the sparsity level, obeying (iv) ahead; (ii) the class of functions $\mathcal{F}_j = \{w \mapsto \psi_j\{w, \alpha, \widetilde{h}(z_j)\} : \alpha \in \mathcal{A}_j, \widetilde{h} \in \mathcal{H}_j \cup \{h_j\}\}$ is pointwise measurable and obeys the entropy condition $\mathrm{ent}(\varepsilon, \mathcal{F}_j) \leq C_1 M s \log(a_n/\varepsilon)$ for all $0 < \varepsilon \leq 1$; (iii) the class $\mathcal{F}_j$ has measurable envelope $F_j \geq \sup_{f \in \mathcal{F}_j} |f|$, such that $F = \max_{j=1,\ldots,p_1} F_j$ obeys $E\{F^q(w)\} \leq$*

C_1 for some $q \geq 4$; and (iv) the dimensions p_1, p, and s obey the growth conditions:
$n^{-1/2}\{(s \log a_n)^{1/2} + n^{-1/2+1/q} s \log a_n\} \leq \rho_n, \;\; \rho_n^{\varsigma/2}(L_{2n} s \log a_n)^{1/2} + n^{1/2} L_{1n} \rho_n^2 \leq \delta_n b_n^{-1}$.

Condition 3 (i) requires reasonable behavior of sparse estimators $\widehat{h}_j$. In the previous section, this type of behavior occurred in the cases where h_j consisted of a part of a median regression function and a conditional expectation function in an auxiliary equation. There are many conditions in the literature that imply these conditions from primitive assumptions. For the case with $q = \infty$, Condition 3 (vi) implies the following restrictions on the sparsity indices: $(s^2 \log^3 a_n)/n \to 0$ for the case where $\varsigma = 2$, which typically happens when ψ_j is smooth, and $(s^3 \log^5 a_n)/n \to 0$ for the case where $\varsigma = 1$, which typically happens when ψ_j is non-smooth. Condition 3 (iii) bounds the moments of the envelopes, and it can be relaxed to a bound that grows with n, with an appropriate strengthening of the growth conditions stated in (iv).

Condition 3 (ii) implicitly requires ψ_j not to increase entropy too much; it holds, for example, when ψ_j is a monotone transformation, as in the case of median regression, or a Lipschitz transformation; see [32]. The entropy bound is formulated in terms of the upper bound s on the sparsity of the estimators and p the dimension of the overall approximating model appearing via a_n. In principle our main result below applies to non-sparse estimators as well, as long as the entropy bound specified in Condition 3 (ii) holds, with index (s, p) interpreted as measures of effective complexity of the relevant function classes.

Recall that $\Gamma_j = \partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\}]$; see Condition 2 (iii). Define

$$\sigma_j^2 = E[\Gamma_j^{-2} \psi_j^2\{w, \alpha_j, h_j(z_j)\}], \quad \phi_j(w) = -\sigma_j^{-1} \Gamma_j^{-1} \psi_j\{w, \alpha_j, h_j(z_j)\} \quad (j = 1, \ldots, p_1).$$

The following is the main theorem of this section; its proof is found in Appendix A.

Theorem 2. *Under Conditions 2 and 3, uniformly in $P \in \mathcal{P}_n$, with probability $1 - o(1)$,*

$$\max_{j=1,\ldots,p_1} \left| n^{1/2} \sigma_j^{-1} (\widehat{\alpha}_j - \alpha_j) - n^{-1/2} \sum_{i=1}^n \phi_j(w_i) \right| = o(b_n^{-1}), \quad n \to \infty.$$

An immediate implication is a corollary on the asymptotic normality uniform in $P \in \mathcal{P}_n$ and $j = 1, \ldots, p_1$, which follows from Lyapunov's central limit theorem for triangular arrays.

Corollary 2. *Under the conditions of Theorem 2,*

$$\max_{j=1,\ldots,p_1} \sup_{P \in \mathcal{P}_n} \sup_{t \in \mathbb{R}} \left| \mathrm{pr}_P \left\{ n^{1/2} \sigma_j^{-1} (\widehat{\alpha}_j - \alpha_j) \leq t \right\} - \Phi(t) \right| = o(1), \quad n \to \infty.$$

This implies, provided $\max_{j=1,\ldots,p_1} |\widehat{\sigma}_j - \sigma_j| = o_P(1)$ uniformly in $P \in \mathcal{P}_n$, that

$$\max_{j=1,\ldots,p_1} \sup_{P \in \mathcal{P}_n} \left| \mathrm{pr}_P \left\{ \alpha_j \in [\widehat{\alpha}_j \pm \widehat{\sigma}_j n^{-1/2} \Phi^{-1}(1 - \xi/2)] \right\} - (1 - \xi) \right| = o(1), \quad n \to \infty.$$

This result leads to marginal confidence intervals for α_j, and shows that they are valid uniformly in $P \in \mathcal{P}_n$ and $j = 1, \ldots, p_1$.

Another useful implication is the high-dimensional central limit theorem uniformly over rectangles in $\mathbb{R}^{p_1}$, provided that $(\log p_1)^7 = o(n)$, which follows from Corollary 2.1 in [9]. Let $\mathcal{N} = (\mathcal{N}_j)_{j=1}^p$ be a normal random vector in $\mathbb{R}^{p_1}$ with mean zero and covariance matrix $[E\{\phi_j(w)\phi_{j'}(w)\}]_{j,j'=1}^{p_1}$. Let $\mathcal{R}$ be a collection of rectangles R in $\mathbb{R}^{p_1}$ of the form

$$R = \left\{ z \in \mathbb{R}^{p_1} : \max_{j \in A} z_j \leq t, \max_{j \in B} (-z_j) \leq t \right\} \quad (t \in \mathbb{R}, A, B \subset \{1, \ldots, p_1\}).$$

For example, when $A = B = \{1, \ldots, p_1\}$, $R = \{z \in \mathbb{R}^{p_1} : \max_{j=1,\ldots,p_1} |z_j| \leq t\}$.

Corollary 3. *Under the conditions of Theorem 2, provided that* $(\log p_1)^7 = o(n)$,

$$\sup_{P \in \mathcal{P}_n} \sup_{R \in \mathcal{R}} \left| \mathrm{pr}_P \left[n^{1/2} \{\sigma_j^{-1}(\widehat{\alpha}_j - \alpha_j)\}_{j=1}^{p_1} \in R \right] - \mathrm{pr}_P(\mathcal{N} \in R) \right| = o(1), \quad n \to \infty.$$

This implies, in particular, that for $c_{1-\xi} = (1-\xi)$*-quantile of* $\max_{j=1,\ldots,p_1} |\mathcal{N}_j|$,

$$\sup_{P \in \mathcal{P}_n} \left| \mathrm{pr}_P \left(\alpha_j \in [\widehat{\alpha}_j \pm c_{1-\xi} \sigma_j n^{-1/2}],\ j = 1, \ldots, p_1 \right) - (1-\xi) \right| = o(1), \quad n \to \infty.$$

This result leads to simultaneous confidence bands for $(\alpha_j)_{j=1}^{p_1}$ that are valid uniformly in $P \in \mathcal{P}_n$. Moreover, Corollary 3 is immediately useful for testing multiple hypotheses about $(\alpha_j)_{j=1}^{p_1}$ via the step-down methods of [28] which control the family-wise error rate; see [9] for further discussion of multiple testing with $p_1 \gg n$.

In practice the distribution of $\mathcal{N}$ is unknown, since its covariance matrix is unknown, but it can be approximated by the Gaussian multiplier bootstrap, which generates a vector

$$\mathcal{N}^* = (\mathcal{N}_j^*)_{j=1}^{p_1} = \left\{ \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i \widehat{\phi}_j(w_i) \right\}_{j=1}^{p_1}, \tag{19}$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables, independent of the data $(w_i)_{i=1}^n$, and $\widehat{\phi}_j$ are any estimators of ϕ_j, such that

$$\max_{j,j' \in \{1,\ldots,p_1\}} |E_n\{\widehat{\phi}_j(w)\widehat{\phi}_{j'}(w)\} - E_n\{\phi_j(w)\phi_{j'}(w)\}| = o_P(b_n^{-4})$$

uniformly in $P \in \mathcal{P}_n$. Let $\widehat{\sigma}_j^2 = E_n\{\widehat{\phi}_j^2(w)\}$. Theorem 3.2 in [9] then implies the following result.

Corollary 4. *Under the conditions of Theorem 2, provided that* $(\log p_1)^7 = o(n)$*, with probability* $1 - o(1)$ *uniformly in* $P \in \mathcal{P}_n$*,*

$$\sup_{P \in \mathcal{P}_n} \sup_{R \in \mathcal{R}} |\mathrm{pr}_P\{\mathcal{N}^* \in R \mid (w_i)_{i=1}^n\} - \mathrm{pr}_P(\mathcal{N} \in R)| = o(1).$$

This implies, in particular, that for $\widehat{c}_{1-\xi} = (1-\xi)$*-conditional quantile of* $\max_{j=1,\ldots,p_1} |\mathcal{N}_j^*|$,

$$\sup_{P \in \mathcal{P}_n} \left| \mathrm{pr}_P \left(\alpha_j \in [\widehat{\alpha}_j \pm \widehat{c}_{1-\xi} \widehat{\sigma}_j n^{-1/2}],\ j = 1, \ldots, p_1 \right) - (1-\xi) \right| = o(1).$$

Comment 3.1. The proof of Theorem 2 shows that the orthogonality condition (18) can be replaced by a more general orthogonality condition:

$$E[\eta(z_j)^{\mathrm{T}}\{\widetilde{h}_j(z_j) - h_j(z_j)\}] = 0, \quad (\widetilde{h}_j \in \mathcal{H}_j,\ j = 1, \ldots, p_1), \tag{20}$$

where $\eta(z_j) = \partial_t E\{\psi_j(w, \alpha_j, t) \mid z_j\}|_{t=h_j(z_j)}$, or even more general condition of approximate orthogonality: $E[\eta(z_j)^{\mathrm{T}}\{\widetilde{h}_j(z_j) - h_j(z_j)\}] = o(n^{-1/2}b_n^{-1})$ uniformly in $\widetilde{h}_j \in \mathcal{H}_j$ and $j = 1, \ldots, p_1$. The generalization (20) has a number of benefits, which could be well illustrated by the median regression model of Section 1, where the conditional moment restriction $E(v_i \mid x_i) = 0$ could be now replaced by the unconditional one $E(v_i x_i) = 0$, which allows for more general forms of data-generating processes.

4. MONTE CARLO EXPERIMENTS

We consider the regression model

$$y_i = d_i\alpha_0 + x_i^{\mathrm{T}}(c_y\theta_0) + \epsilon_i, \quad d_i = x_i^{\mathrm{T}}(c_d\theta_0) + v_i, \tag{21}$$

where $\alpha_0 = 1/2$, $\theta_{0j} = 1/j^2$ $(j = 1, \ldots, 10)$, and $\theta_{0j} = 0$ otherwise, $x_i = (1, z_i^{\mathrm{T}})^{\mathrm{T}}$ consists of an intercept and covariates $z_i \sim N(0, \Sigma)$, and the errors ϵ_i and v_i are independently and identically distributed as $N(0, 1)$. The dimension p of the controls x_i is 300, and the sample size n is 250. The covariance matrix Σ has entries $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0{\cdot}5$. The coefficients c_y and c_d determine the R^2 in the equations $y_i - d_i\alpha_0 = x_i^{\mathrm{T}}(c_y\theta_0) + \epsilon_i$ and $d_i = x_i^{\mathrm{T}}(c_d\theta_0) + v_i$. We vary the R^2 in the two equations, denoted by R_y^2 and R_d^2 respectively, in the set $\{0, 0{\cdot}1, \ldots, 0{\cdot}9\}$, which results in 100 different designs induced by the different pairs of (R_y^2, R_d^2); we performed 500 Monte Carlo repetitions for each.

The first equation in (32) is a sparse model. However, unless c_y is very large, the decay of the components of θ_0 rules out the typical assumption that the coefficients of important regressors are well separated from zero. Thus we anticipate that the standard post-selection inference procedure, discussed around (3), would work poorly in the simulations. In contrast, from the prior theoretical arguments, we anticipate that our instrumental median estimator would work well.

The simulation study focuses on Algorithm 1, since Algorithm 2 performs similarly. Standard errors are computed using (11). As the main benchmark we consider the standard post-model selection estimator $\widetilde{\alpha}$ based on the post ℓ_1-penalized median regression method (3).

In Figure 1, we display the empirical false rejection probability of tests of a true hypothesis $\alpha = \alpha_0$, with nominal size 5%. The false rejection probability of the standard post-model selection inference procedure based upon $\widetilde{\alpha}$ deviates sharply from the nominal size. This confirms the anticipated failure, or lack of uniform validity, of inference based upon the standard post-model selection procedure in designs where coefficients are not well separated from zero so that perfect model selection does not happen. In sharp contrast, both of our proposed procedures, based on estimator $\check{\alpha}$ and the result (10) and on the statistic L_n and the result (13), closely track the nominal size. This is achieved uniformly over all the designs considered in the study, and confirms the theoretical results of Corollary 1.

In Figure 2, we compare the performance of the standard post-selection estimator $\widetilde{\alpha}$ and our proposed post-selection estimator $\check{\alpha}$. We use three different measures of performance of the two approaches: mean bias, standard deviation, and root mean square error. The significant bias for the standard post-selection procedure occurs when the main regressor d_i is correlated with other controls x_i. The proposed post-selection estimator $\check{\alpha}$ performs well in all three measures. The root mean square errors of $\check{\alpha}$ are typically much smaller than those of $\widetilde{\alpha}$, fully consistent with our theoretical results and the semiparametric efficiency of $\check{\alpha}$.

SUPPLEMENTARY MATERIAL

In the supplementary material we provide omitted proofs, technical lemmas, discuss extensions to the heteroscedastic case, and alternative implementations.

APPENDIX A. PROOF OF THEOREM 2

A.1. **A maximal inequality.** We first state a maximal inequality used in the proof of Theorem 2.

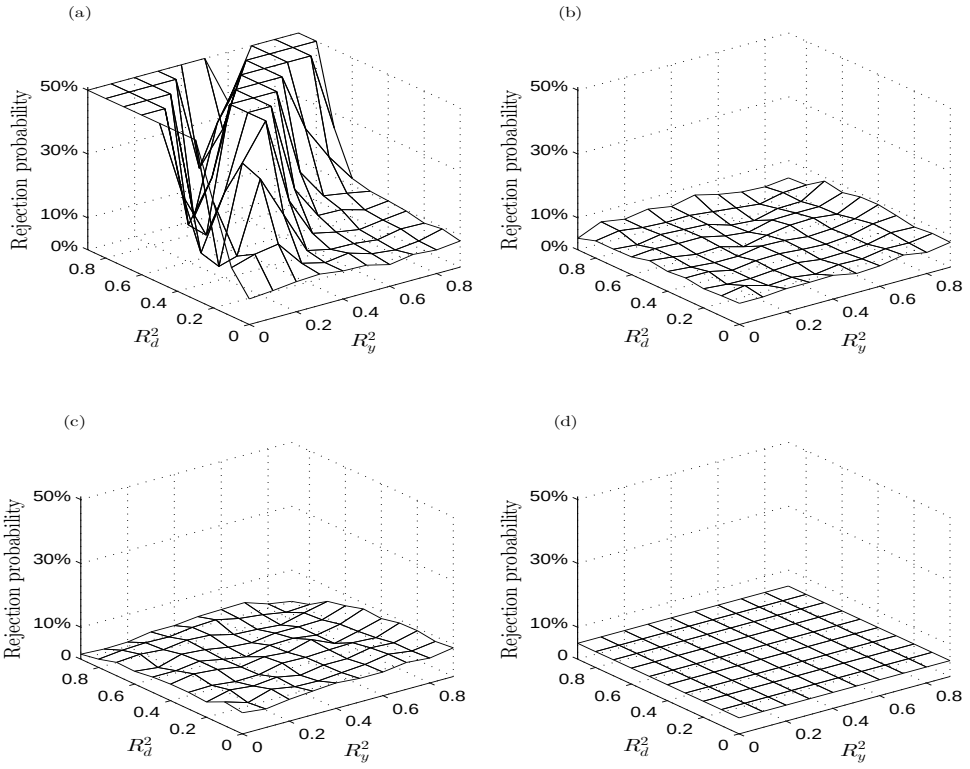

FIGURE 1. The empirical false rejection probabilities of the nominal 5% level tests based on: (a) the standard post-model selection procedure based on $\widetilde{\alpha}$, (b) the proposed post-model selection procedure based on $\check{\alpha}$, (c) the score statistic L_n, and (d) an ideal procedure with the false rejection rate equal to the nominal size.

Lemma 1. *Let $w, w_1, \ldots, w_n$ be independent and identically distributed random variables taking values in a measurable space, and let $\mathcal{F}$ be a pointwise measurable class of functions on that space. Suppose that there is a measurable envelope $F \geq \sup_{f\in\mathcal{F}} |f|$ such that $E\{F^q(w)\} < \infty$ for some $q \geq 2$. Consider the empirical process indexed by $\mathcal{F}$: $G_n(f) = n^{-1/2}\sum_{i=1}^n [f(w_i) - E\{f(w)\}], f \in \mathcal{F}$. Let $\sigma > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} E\{f^2(w)\} \leq \sigma^2 \leq E\{F^2(w)\}$. Moreover, suppose that there exist constants $A \geq e$ and $s \geq 1$ such that $\mathrm{ent}(\varepsilon, \mathcal{F}) \leq s\log(A/\varepsilon)$ for all $0 < \varepsilon \leq 1$. Then*

$$E\left\{\sup_{f\in\mathcal{F}} |G_n(f)|\right\} \leq K\Bigg[\left\{s\sigma^2 \log(A[E\{F^2(w)\}]^{1/2}/\sigma)\right\}^{1/2} + n^{-1/2+1/q} s[E\{F^q(w)\}]^{1/q} \log(A[E\{F^2(w)\}]^{1/2}/\sigma)\Bigg],$$

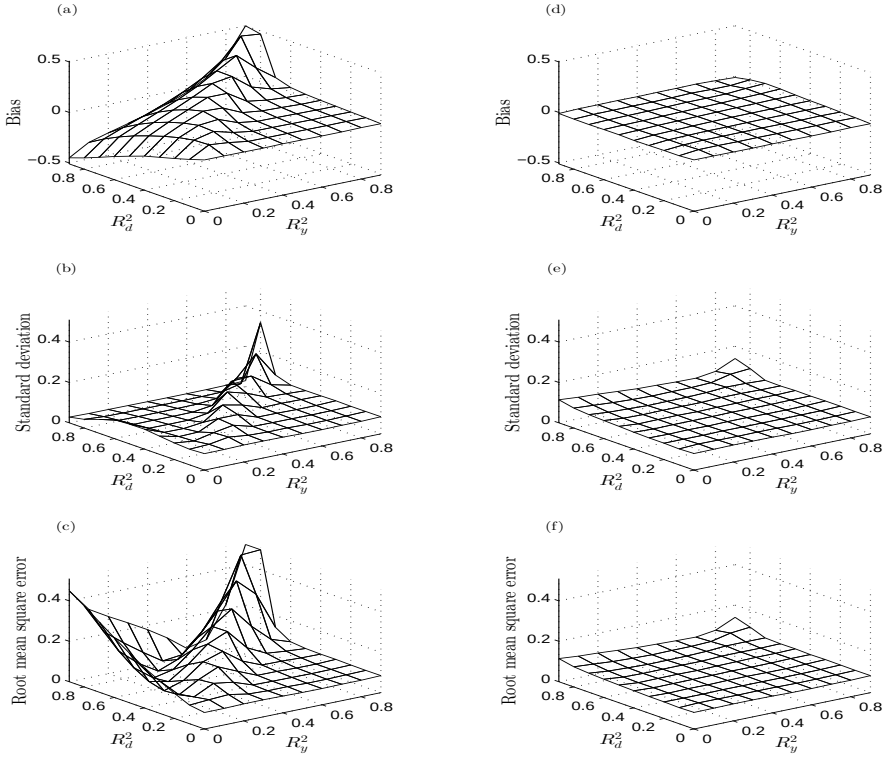

FIGURE 2. Mean bias (top row), standard deviation (middle row), root mean square (bottom row) of the standard post-model selection estimator $\widetilde{\alpha}$ (panels (a)-(c)), and of the proposed post-model selection estimator $\check{\alpha}$ (panels (d)-(f)).

where K is a universal constant. Moreover, for every $t \geq 1$, with probability not less than $1 - t^{-q/2}$,

$$\sup_{f \in \mathcal{F}} |G_n(f)| \leq 2E \left\{ \sup_{f \in \mathcal{F}} |G_n(f)| \right\} + K_q \left(\sigma t^{1/2} + n^{-1/2+1/q} [E\{F^q(w)\}]^{1/q} t \right),$$

where K_q is a constant that depends only on q.

Proof. The first and second inequalities follow from Corollary 5.1 and Theorem 5.1 in [10] applied with $\alpha = 1$, using that $[E\{\max_{i=1,\ldots,n} F^2(w_i)\}]^{1/2} \leq [E\{\max_{i=1,\ldots,n} F^q(w_i)\}]^{1/q} \leq n^{1/q}[E\{F^q(w)\}]^{1/q}$. □

A.2. **Proof of Theorem 2.** It suffices to prove the theorem under any sequence $P = P_n \in \mathcal{P}_n$. We shall suppress the dependence of P on n in the proof. In this proof, let C denote a generic positive constant that may differ in each appearance, but that does not depend on the sequence $P \in \mathcal{P}_n$, n, or $j = 1, \ldots, p_1$. Recall that the sequence $\rho_n \downarrow 0$ satisfies the growth conditions in

Condition 3 (iv). We divide the proof into three steps. Below we use the following notation: for any given function $g : \mathcal{W} \to \mathbb{R}$, $G_n(g) = n^{-1/2} \sum_{i=1}^n [g(w_i) - E\{g(w)\}]$.

Step 1. Let $\widetilde{\alpha}_j$ be any estimator such that with probability $1 - o(1)$, $\max_{j=1,\ldots,p_1} |\widetilde{\alpha}_j - \alpha_j| \leq C\rho_n$. We wish to show that, with probability $1 - o(1)$,

$$E_n[\psi_j\{w, \widetilde{\alpha}_j, \widehat{h}_j(z_j)\}] = E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + \Gamma_j(\widetilde{\alpha}_j - \alpha_j) + o(n^{-1/2} b_n^{-1}),$$

uniformly in $j = 1, \ldots, p_1$. Expand

$$\begin{aligned} E_n[\psi_j\{w, \widetilde{\alpha}_j, \widehat{h}_j(z_j)\}] &= E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + E[\psi_j\{w, \alpha, \widetilde{h}(z_j)\}]|_{\alpha=\widetilde{\alpha}_j, \widetilde{h}=\widehat{h}_j} \\ &\quad + n^{-1/2} G_n[\psi_j\{w, \widetilde{\alpha}_j, \widehat{h}_j(z_j)\} - \psi_j\{w, \alpha_j, h_j(z_j)\}] = I_j + II_j + III_j, \end{aligned}$$

where we have used $E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = 0$. We first bound III_j. Observe that, with probability $1 - o(1)$, $\max_{j=1,\ldots,p_1} |III_j| \leq n^{-1/2} \sup_{f \in \mathcal{F}} |G_n(f)|$, where $\mathcal{F}$ is the class of functions defined by

$$\mathcal{F} = \{w \mapsto \psi_j\{w, \alpha, \widetilde{h}(z_j)\} - \psi_j\{w, \alpha_j, h_j(z_j)\} : j = 1, \ldots, p_1, \widetilde{h} \in \mathcal{H}_j, \alpha \in \mathcal{A}_j, |\alpha - \alpha_j| \leq C\rho_n\},$$

which has $2F$ as an envelope. We apply Lemma 1 to this class of functions. By Condition 3 (ii) and a simple covering number calculation, we have $\mathrm{ent}(\varepsilon, \mathcal{F}) \leq Cs \log(a_n/\varepsilon)$. By Condition 2 (ii), $\sup_{f \in \mathcal{F}} E\{f^2(w)\}$ is bounded by

$$\sup_{\substack{j=1,\ldots,p_1, (\alpha, \widetilde{h}) \in \mathcal{A}_j \times \mathcal{H}_j \\ |\alpha - \alpha_j| \leq C\rho_n}} E\left\{E\left(\left[\psi_j\{w, \alpha, \widetilde{h}(z_j)\} - \psi_j\{w, \alpha_j, h_j(z_j)\}\right]^2 \mid z_j\right)\right\} \leq C L_{2n} \rho_n^{\varsigma},$$

where we have used the fact that $E[\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}^2] \leq C\rho_n^2$ for all $m = 1, \ldots, M$ whenever $\widetilde{h} = (\widetilde{h}_m)_{m=1}^M \in \mathcal{H}_j$. Hence applying Lemma 1 with $t = \log n$, we conclude that, with probability $1 - o(1)$,

$$n^{1/2} \max_{j=1,\ldots,p_1} |III_j| \leq \sup_{f \in \mathcal{F}} |G_n(f)| \leq C\{\rho_n^{\varsigma/2} (L_{2n} s \log a_n)^{1/2} + n^{-1/2+1/q} s \log a_n\} = o(b_n^{-1}),$$

where the last equality follows from Condition 3 (iv).

Next, we expand II_j. Pick any $\alpha \in \mathcal{A}_j$ with $|\alpha - \alpha_j| \leq C\rho_n$, $\widetilde{h} = (\widetilde{h}_m)_{m=1}^M \in \mathcal{H}_j$. Then by Taylor's theorem, for any $j = 1, \ldots, p_1$ and $z_j \in \mathcal{Z}_j$, there exists a vector $(\bar{\alpha}(z_j), \bar{t}(z_j)^{\mathrm{T}})^{\mathrm{T}}$ on the line segment joining $(\alpha, \widetilde{h}(z_j)^{\mathrm{T}})^{\mathrm{T}}$ and $(\alpha_j, h_j(z_j)^{\mathrm{T}})^{\mathrm{T}}$ such that $E[\psi_j\{w, \alpha, \widetilde{h}(z_j)\}]$ can be written as

$$\begin{aligned} &E[\psi_j\{w, \alpha_j, h_j(z_j)\}] + E(\partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j])(\alpha - \alpha_j) \\ &+ \textstyle\sum_{m=1}^M E\{E\left(\partial_{t_m} E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j]\right)\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}\} \\ &+ 2^{-1} E(\partial_\alpha^2 E[\psi_j\{w, \bar{\alpha}(z_j), \bar{t}(z_j)\} \mid z_j])(\alpha - \alpha_j)^2 \\ &+ 2^{-1} \textstyle\sum_{m,m'=1}^M E(\partial_{t_m} \partial_{t_{m'}} E[\psi_j\{w, \bar{\alpha}(z_j), \bar{t}(z_j)\} \mid z_j]\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}\{\widetilde{h}_{m'}(z_j) - h_{jm'}(z_j)\}) \\ &+ \textstyle\sum_{m=1}^M E(\partial_\alpha \partial_{t_m} E[\psi_j\{w, \bar{\alpha}(z_j), \bar{t}(z_j)\} \mid z_j](\alpha - \alpha_j)\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}). \end{aligned} \tag{22}$$

The third term is zero because of the orthogonality condition (18). Condition 2 (ii) guarantees that the expectation and derivative can be interchanged for the second term, that is, $E\left(\partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j]\right) = \partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = \Gamma_j$. Moreover, by the same

condition, each of the last three terms is bounded by $CL_{1n}\rho_n^2 = o(n^{-1/2}b_n^{-1})$, uniformly in $j = 1, \ldots, p_1$. Therefore, with probability $1 - o(1)$, $II_j = \Gamma_j(\widetilde{\alpha}_j - \alpha_j) + o(n^{-1/2}b_n^{-1})$, uniformly in $j = 1, \ldots, p_1$. Combining the previous bound on III_j with these bounds leads to the desired assertion.

Step 2. We wish to show that with probability $1 - o(1)$, $\inf_{\alpha \in \widehat{\mathcal{A}}_j} |E_n[\psi_j\{w, \alpha, \widehat{h}_j(z_j)\}]| = o(n^{-1/2}b_n^{-1})$, uniformly in $j = 1, \ldots, p_1$. Define $\alpha_j^* = \alpha_j - \Gamma_j^{-1}E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}]$ $(j = 1, \ldots, p_1)$. Then we have $\max_{j=1,\ldots,p_1} |\alpha_j^* - \alpha_j| \leq C \max_{j=1,\ldots,p_1} |E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}]|$. Consider the class of functions $\mathcal{F}' = \{w \mapsto \psi_j\{w, \alpha_j, h_j(z_j)\} : j = 1, \ldots, p_1\}$, which has F as an envelope. Since this class is finite with cardinality p_1, we have $\mathrm{ent}(\varepsilon, \mathcal{F}') \leq \log(p_1/\varepsilon)$. Hence applying Lemma 1 to $\mathcal{F}'$ with $\sigma = [E\{F^2(w)\}]^{1/2} \leq C$ and $t = \log n$, we conclude that with probability $1 - o(1)$,

$$\max_{j=1,\ldots,p_1} |E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}]| \leq Cn^{-1/2}\{(\log a_n)^{1/2} + n^{-1/2+1/q} \log a_n\} \leq Cn^{-1/2} \log a_n.$$

Since $\widehat{\mathcal{A}}_j \supset [\alpha_j \pm c_1 n^{-1/2} \log^2 a_n]$ with probability $1 - o(1)$, $\alpha_j^* \in \widehat{\mathcal{A}}_j$ with probability $1 - o(1)$.

Therefore, using Step 1 with $\widetilde{\alpha}_j = \alpha_j^*$, we have, with probability $1 - o(1)$,

$$\inf_{\alpha \in \widehat{\mathcal{A}}_j} |E_n[\psi_j\{w, \alpha, \widehat{h}_j(z_j)\}]| \leq |E_n[\psi_j\{w, \alpha_j^*, \widehat{h}_j(z_j)\}]| = o(n^{-1/2}b_n^{-1}),$$

uniformly in $j = 1, \ldots, p_1$, where we have used the fact that $E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + \Gamma_j(\alpha_j^* - \alpha_j) = 0$.

Step 3. We wish to show that with probability $1 - o(1)$, $\max_{j=1,\ldots,p_1} |\widehat{\alpha}_j - \alpha_j| \leq C\rho_n$. By Step 2 and the definition of $\widehat{\alpha}_j$, with probability $1 - o(1)$, we have $\max_{j=1,\ldots,p_1} |E_n[\psi_j\{w, \widehat{\alpha}_j, \widehat{h}_j(z_j)\}]| = o(n^{-1/2}b_n^{-1})$. Consider the class of functions $\mathcal{F}'' = \{w \mapsto \psi_j\{w, \alpha, \widetilde{h}(z_j)\} : j = 1, \ldots, p_1, \alpha \in \mathcal{A}_j, \widetilde{h} \in \mathcal{H}_j \cup \{h_j\}\}$. Then with probability $1 - o(1)$,

$$|E_n[\psi_j\{w, \widehat{\alpha}_j, \widehat{h}_j(z_j)\}]| \geq \left|E[\psi_j\{w, \alpha, \widetilde{h}(z_j)\}]|_{\alpha=\widehat{\alpha}_j, \widetilde{h}=\widehat{h}_j}\right| - n^{-1/2} \sup_{f \in \mathcal{F}} |G_n(f)|,$$

uniformly in $j = 1, \ldots, p_1$. Observe that $\mathcal{F}''$ has F as an envelope and, by Condition 3 (ii) and a simple covering number calculation, $\mathrm{ent}(\varepsilon, \mathcal{F}'') \leq Cs \log(a_n/\varepsilon)$. Then applying Lemma 1 with $\sigma = [E\{F^2(w)\}]^{1/2} \leq C$ and $t = \log n$, we have, with probability $1 - o(1)$,

$$n^{-1/2} \sup_{f \in \mathcal{F}''} |G_n(f)| \leq Cn^{-1/2}\{(s \log a_n)^{1/2} + n^{-1/2+1/q} s \log a_n\} = O(\rho_n).$$

Moreover, application of the expansion (22) with $\alpha_j = \alpha$ together with the Cauchy–Schwarz inequality implies that $|E[\psi_j\{w, \alpha, \widetilde{h}(z_j)\}] - E[\psi_j\{w, \alpha, h_j(z_j)\}]|$ is bounded by $C(\rho_n + L_{1n}\rho_n^2) = O(\rho_n)$, so that with probability $1 - o(1)$,

$$\left|E[\psi_j\{w, \alpha, \widetilde{h}(z_j)\}]|_{\alpha=\widehat{\alpha}_j, \widetilde{h}=\widehat{h}_j}\right| \geq \left|E[\psi_j\{w, \alpha, h_j(z_j)\}]|_{\alpha=\widehat{\alpha}_j}\right| - O(\rho_n),$$

uniformly in $j = 1, \ldots, p_1$, where we have used Condition 2 (ii) together with the fact that $E[\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}^2] \leq C\rho_n^2$ for all $m = 1, \ldots, M$ whenever $\widetilde{h} = (\widetilde{h}_m)_{m=1}^M \in \mathcal{H}_j$. By Condition 2 (iv), the first term on the right side is bounded from below by $(1/2)\{|\Gamma_j(\widehat{\alpha}_j - \alpha_j)| \wedge c_1\}$, which, combined with the fact that $|\Gamma_j| \geq c_1$, implies that with probability $1 - o(1)$, $|\widehat{\alpha}_j - \alpha_j| \leq o(n^{-1/2}b_n^{-1}) + O(\rho_n) = O(\rho_n)$, uniformly in $j = 1, \ldots, p_1$.

Step 4. By Steps 1 and 3, with probability $1 - o(1)$,

$$E_n[\psi_j\{w, \widehat{\alpha}_j, \widehat{h}_j(z_j)\}] = E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + \Gamma_j(\widehat{\alpha}_j - \alpha_j) + o(n^{-1/2}b_n^{-1}),$$

uniformly in $j = 1, \ldots, p_1$. Moreover, by Step 2, with probability $1 - o(1)$, the left side is $o(n^{-1/2}b_n^{-1})$ uniformly in $j = 1, \ldots, p_1$. Solving this equation with respect to $(\widehat{\alpha}_j - \alpha_j)$ leads to the conclusion of the theorem. □

REFERENCES

[1] Donald WK Andrews. Empirical process methods in econometrics. *Handbook of Econometrics*, 4:2247–2294, 1994.

[2] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2430, November 2012.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Ann. Statist.*, 39(1):82–130, 2011.

[4] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics: The 2010 World Congress of the Econometric Society*, 3:245–295, 2013.

[5] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.*, 81:608–650, 2014.

[6] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42:757–788, 2014.

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

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

[9] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

[10] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42:1564–1597, 2014.

[11] Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *J. Econometrics*, 142:379–398, 2008.

[12] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized Processes: Limit Theory and Statistical Applications*. Springer, New York, 2009.

[13] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *J. Multivariate Anal.*, 73(1):120–135, 2000.

[14] P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, 1:799–821, 1973.

[15] Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.*, 86(1):4–29, 2004.

[16] Roger Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

[17] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

[18] Sokbae Lee. Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory*, 19:1–31, 2003.

[19] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Econometric Theory*, 21:21–59, 2005.

[20] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

[21] Hua Liang, Suojin Wang, James M. Robins, and Raymond J. Carroll. Estimation in partially linear models with missing covariates. *J. Amer. Statist. Assoc.*, 99(466):357–367, 2004.

[22] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[23] S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.*, 12:1298–1309, 1984.
[24] S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.*, 13:1251–1638, 1985.
[25] J. L. Powell. Censored regression quantiles. *J. Econometrics*, 32:143–155, 1986.
[26] James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, 90(429):122–129, 1995.
[27] P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.
[28] Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, July 2005.
[29] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, 59:3434–3447, 2013.
[30] R. J. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.
[31] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.
[32] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, 1996.
[33] Lie Wang. L_1 penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.*, 120:135–151, 2013.
[34] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242, 2014.

Suplementary Material
Uniform Post Selection Inference for Least Absolute Deviation Regression and Other Z-estimation Problems

This supplementary material contains omitted proofs, technical lemmas, discussion of the extension to the heteroscedastic case, and alternative implementations of the estimator.

APPENDIX B. ADDITIONAL NOTATION IN THE SUPPLEMENTARY MATERIAL

In addition to the notation used in the main text, we will use the following notation. Denote by $\|\cdot\|_\infty$ the maximal absolute element of a vector. Given a vector $\delta \in \mathbb{R}^p$ and a set of indices $T \subset \{1,\ldots,p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector such that $(\delta_T)_j = \delta_j$ if $j \in T$ and $(\delta_T)_j = 0$ if $j \notin T$. For a sequence $(z_i)_{i=1}^n$ of constants, we write $\|z_i\|_{2,n} = \{E_n(z_i^2)\}^{1/2} = (n^{-1}\sum_{i=1}^n z_i^2)^{1/2}$. For example, for a vector $\delta \in \mathbb{R}^p$ and p-dimensional regressors $(x_i)_{i=1}^n$, $\|x_i^T\delta\|_{2,n} = [E_n\{(x_i^{\mathrm{T}}\delta)^2\}]^{1/2}$ denotes the empirical prediction norm of δ. Denote by $\|\cdot\|_{P,2}$ the population L^2-seminorm. We also 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 $a = O_P(b)$.

APPENDIX C. GENERALIZATION AND ADDITIONAL RESULTS FOR THE LEAST ABSOLUTE DEVIATION MODEL

C.1. **Generalization of Section 2 to heteroscedastic case.** We emphasize that both proposed algorithms exploit the homoscedasticity of the model (1) with respect to the error term ϵ_i. The generalization to the heteroscedastic case can be achieved as follows. Recall the model $y_i = d_i\alpha_0 + x_i^{\mathrm{T}}\beta_0 + \epsilon_i$ where ϵ_i is now not necessarily independent of d_i and x_i but obeys the conditional median restriction $\mathrm{pr}(\epsilon_i \leq 0 \mid d_i, x_i) = 1/2$. To achieve the semiparametric efficiency bound in this general case, we need to consider the weighted version of the auxiliary equation (4). Specifically, we rely on the weighted decomposition:

$$f_i d_i = f_i x_i^{\mathrm{T}}\theta_0^* + v_i^*,\ E(f_i v_i^* \mid x_i) = 0 \quad (i = 1,\ldots,n), \tag{23}$$

where the weights are the conditional densities of the error terms ϵ_i evaluated at their conditional medians of zero:

$$f_i = f_{\epsilon_i}(0 \mid d_i, x_i) \quad (i = 1,\ldots,n), \tag{24}$$

which in general vary under heteroscedasticity. With that in mind it is straightforward to adapt the proposed algorithms when the weights $(f_i)_{i=1}^n$ are known. For example Algorithm 1 becomes as follows.

Algorithm 1′. *The algorithm is based on post-model selection estimators.*
Step *(i). Run post-ℓ_1-penalized median regression of y_i on d_i and x_i; keep fitted value $x_i^{\mathrm{T}}\widetilde{\beta}$.*
Step *(ii). Run the post-lasso estimator of $f_i d_i$ on $f_i x_i$; keep the residual $\widehat{v}_i^* = f_i(d_i - x_i^{\mathrm{T}}\widetilde{\theta})$.*
Step *(iii). Run instrumental median regression of $y_i - x_i^{\mathrm{T}}\widetilde{\beta}$ on d_i using $\widehat{v}_i^*$ as the instrument. Report $\check{\alpha}$ and/or perform inference.*

Analogously, we obtain Algorithm $2'$, as a generalization of Algorithm 2 in the main text, based on regularized estimators, by removing the word "post" in Algorithm $1'$.

Under similar regularity conditions, uniformly over a large collection $\mathcal{P}_n^*$ of distributions of $\{(y_i, d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$, the estimator $\check{\alpha}$ above obeys

$$\{4E(v_i^{*2})\}^{1/2} n^{1/2}(\check{\alpha} - \alpha_0) \to N(0,1)$$

in distribution. Moreover, the criterion function at the true value α_0 in Step (iii) also has a pivotal behavior, namely

$$nL_n(\alpha_0) \to \chi_1^2$$

in distribution, which can also be used to construct a confidence region $\widehat{A}_\xi$ based on the L_n-statistic as in (13) with coverage $1-\xi$ uniformly in a suitable collection of distributions.

In practice the density function values $(f_i)_{i=1}^n$ are unknown and need to be replaced by estimates $(\widehat{f}_i)_{i=1}^n$. The analysis of the impact of such estimation is very delicate and is developed in the companion work "Robust inference in high-dimensional approximately sparse quantile regression models" (arXiv:1312.7186), which considers the more general problem of uniformly valid inference for quantile regression models in approximately sparse models.

C.2. **Minimax Efficiency.** The asymptotic variance, $(1/4)\{E(v_i^{*2})\}^{-1}$, of the estimator $\check{\alpha}$ is the semiparametric efficiency bound for estimation of α_0. To see this, given a law P_n with $\|\beta_0\|_0 \vee \|\theta_0^*\|_0 \leq s/2$, we first consider a submodel $\mathcal{P}_n^{\text{sub}} \subset \mathcal{P}_n^*$ such that $P_n \in \mathcal{P}_n^{\text{sub}}$, indexed by the parameter $t = (t_1, t_2) \in \mathbb{R}^2$ for the parametric components α_0, β_0 and described as:

$$\begin{aligned} y_i &= d_i(\alpha_0 + t_1) + x_i^{\mathrm{T}}(\beta_0 + t_2\theta_0^*) + \epsilon_i, \\ f_i d_i &= f_i x_i^{\mathrm{T}}\theta_0^* + v_i^*, \quad E(f_i v_i^* \mid x_i) = 0, \end{aligned}$$

where the conditional density of ϵ_i varies. Here we use $\mathcal{P}_n^*$ to denote the overall model collecting all distributions for which a variant of conditions of Theorem 1 permitting heteroscedasticity is satisfied. In this submodel, setting $t = 0$ leads to the given parametric components α_0, β_0 at P_n. Then by using a similar argument to [18], Section 5, the efficient score for α_0 in this submodel is

$$S_i = 4\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta_0) f_i \{d_i - x_i^{\mathrm{T}}\theta_0^*\} = 4\varphi(\epsilon_i) v_i^*,$$

so that $\{E(S_i^2)\}^{-1} = (1/4)\{E(v_i^{*2})\}^{-1}$ is the efficiency bound at P_n for estimation of α_0 relative to the submodel, and hence relative to the entire model $\mathcal{P}_n^*$, as the bound is attainable by our estimator $\check{\alpha}$ uniformly in P_n in $\mathcal{P}_n^*$. This efficiency bound continues to apply in the homoscedastic model with $f_i = f_\epsilon$ for all i.

C.3. **Alternative implementation via double selection.** An alternative proposal for the method is reminiscent of the double selection method proposed in [5] for partial linear models. This version replaces Step (iii) with a median regression of y on d and all covariates selected in Steps (i) and (ii), that is, the union of the selected sets. The method is described as follows:

Algorithm 3. *The algorithm is based on double selection.*
Step *(i). Run ℓ_1-penalized median regression of y_i on d_i and x_i:*

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} E_n(|y_i - d_i\alpha - x_i^{\mathrm{T}}\beta|) + \frac{\lambda_1}{n}\|\Psi(\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}\|_1.$$

Step *(ii). Run lasso of d_i on x_i:*

$$\widehat{\theta} \in \arg\min_{\theta} E_n\{(d_i - x_i^{\mathrm{T}}\theta)^2\} + \frac{\lambda_2}{n}\|\widehat{\Gamma}\theta\|_1.$$

Step *(iii). Run median regression of y_i on d_i and the covariates selected in Steps (i) and (ii):*

$$(\check{\alpha}, \check{\beta}) \in \arg\min_{\alpha,\beta} \left\{ E_n(|y_i - d_i\alpha - x_i^{\mathrm{T}}\beta|) : \ \mathrm{supp}(\beta) \subset \mathrm{supp}(\widehat{\beta}) \cup \mathrm{supp}(\widehat{\theta}) \right\}.$$

Report $\check{\alpha}$ and/or perform inference.

The double selection algorithm has three main steps: (i) select covariates based on the standard ℓ_1-penalized median regression, (ii) select covariates based on heteroscedastic lasso of the treatment equation, and (ii) run a median regression with the treatment and all selected covariates.

This approach can also be analyzed through Theorem 2 since it creates instruments implicitly. To see that let $\widehat{T}^*$ denote the variables selected in Steps (i) and (ii): $\widehat{T}^* = \mathrm{supp}(\widehat{\beta}) \cup \mathrm{supp}(\widehat{\theta})$. By the first order conditions for $(\check{\alpha}, \check{\beta})$ we have

$$\left\| E_n \left\{ \varphi(y_i - d_i\check{\alpha} - x_i^{\mathrm{T}}\check{\beta})(d_i, x_{i\widehat{T}^*}^{\mathrm{T}})^{\mathrm{T}} \right\} \right\| = O\{(\max_{i=1,\ldots,n} |d_i| + K_n|\widehat{T}^*|^{1/2})(1 + |\widehat{T}^*|)/n\},$$

which creates an orthogonal relation to any linear combination of $(d_i, x_{i\widehat{T}^*}^{\mathrm{T}})^{\mathrm{T}}$. In particular, by taking the linear combination $(d_i, x_{i\widehat{T}^*}^{\mathrm{T}})(1, -\widetilde{\theta}_{\widehat{T}^*}^{\mathrm{T}})^{\mathrm{T}} = d_i - x_{i\widehat{T}^*}^{\mathrm{T}}\widetilde{\theta}_{\widehat{T}^*} = d_i - x_i^{\mathrm{T}}\widetilde{\theta} = \widehat{v}_i$, which is the instrument in Step (ii) of Algorithm 1, we have

$$E_n\{\varphi(y_i - d_i\check{\alpha} - x_i^{\mathrm{T}}\check{\beta})\widehat{z}_i\} = O\{\|(1, -\widetilde{\theta}^{\mathrm{T}})^{\mathrm{T}}\|(\max_{i=1,\ldots,n} |d_i| + K_n|\widehat{T}^*|^{1/2})(1 + |\widehat{T}^*|)/n\}.$$

As soon as the right side is $o_P(n^{-1/2})$, the double selection estimator $\check{\alpha}$ approximately minimizes

$$\widetilde{L}_n(\alpha) = \frac{|E_n\{\varphi(y_i - d_i\alpha - x_i^{\mathrm{T}}\check{\beta})\widehat{v}_i\}|^2}{E_n[\{\varphi(y_i - d_i\check{\alpha} - x_i^{\mathrm{T}}\check{\beta})\}^2\widehat{v}_i^2]},$$

where $\widehat{v}_i$ is the instrument created by Step (ii) of Algorithm 1. Thus the double selection estimator can be seen as an iterated version of the method based on instruments where the Step (i) estimate $\widetilde{\beta}$ is updated with $\check{\beta}$.

APPENDIX D. AUXILIARY RESULTS FOR ℓ_1-PENALIZED MEDIAN REGRESSION AND HETEROSCEDASTIC LASSO

D.1. **Notation.** In this section we state relevant theoretical results on the performance of the estimators: ℓ_1-penalized median regression, post-ℓ_1-penalized median regression, heteroscedastic lasso, and heteroscedastic post-lasso estimators. There results were developed in [3] and [2]. We keep the notation of Sections 1 and 2 in the main text, and let $\widetilde{x}_i = (d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$. Throughout the section, let $c_0 > 1$ be a fixed constant chosen by users. In practice, we suggest to take $c_0 = 1{\cdot}1$ but the analysis is not restricted to this choice. Moreover, let $c_0' = (c_0 + 1)/(c_0 - 1)$. Recall the definition of the minimal and maximal m-sparse eigenvalues of a matrix A as

$$\phi_{\min}(m, A) = \min_{1 \le \|\delta\|_0 \le m} \frac{\delta^{\mathrm{T}} A \delta}{\|\delta\|^2}, \quad \phi_{\max}(m, A) = \max_{1 \le \|\delta\|_0 \le m} \frac{\delta^{\mathrm{T}} A \delta}{\|\delta\|^2},$$

where $m = 1, \ldots, p$. Also recall $\bar{\phi}_{\min}(m) = \phi_{\min}\{m, E(\widetilde{x}_i\widetilde{x}_i^{\mathrm{T}})\}$, $\bar{\phi}_{\max}(m) = \phi_{\max}\{m, E(\widetilde{x}_i\widetilde{x}_i^{\mathrm{T}})\}$, and define $\phi_{\min}(m) = \phi_{\min}\{m, E_n(\widetilde{x}_i\widetilde{x}_i^{\mathrm{T}})\}$, $\phi^x_{\min}(m) = \phi_{\min}\{m, E_n(x_i x_i^{\mathrm{T}})\}$, and $\phi^x_{\max}(m) = \phi_{\max}\{m, E_n(x_i x_i^{\mathrm{T}})\}$. Observe that $\phi_{\max}(m) \leq 2E_n(d^2) + 2\phi^x_{\max}(m)$.

D.2. **ℓ_1-penalized median regression.** Suppose that $\{(y_i, \widetilde{x}_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ are independent and identically distributed random vectors satisfying the conditional median restriction

$$\mathrm{pr}(y_i \leq \widetilde{x}_i^{\mathrm{T}}\eta_0 \mid \widetilde{x}_i) = 1/2 \quad (i = 1, \ldots, n).$$

We consider the estimation of η_0 via the ℓ_1-penalized median regression estimate

$$\widehat{\eta} \in \arg\min_{\eta} E_n(|y_i - \widetilde{x}_i^{\mathrm{T}}\eta|) + \frac{\lambda}{n}\|\Psi\eta\|_1,$$

where $\Psi^2 = \mathrm{diag}\{E_n(\widetilde{x}_{i1}^2), \ldots, E_n(\widetilde{x}_{ip}^2)\}$ is a diagonal matrix of penalty loadings. As established in [3] and [33], under the event that

$$\frac{\lambda}{n} \geq 2c_0\|\Psi^{-1}E_n[\{1/2 - 1(y_i \leq \widetilde{x}_i^{\mathrm{T}}\eta_0)\}\widetilde{x}_i]\|_\infty, \tag{25}$$

the estimator above achieves good theoretical guarantees under mild design conditions. Although η_0 is unknown, we can set λ so that the event in (25) holds with high probability. In particular, the pivotal rule discussed in [3] proposes to set $\lambda = c_0 n\Lambda(1 - \gamma \mid \widetilde{x})$ with $\gamma \to 0$ where

$$\Lambda(1 - \gamma \mid \widetilde{x}) = Q(1 - \gamma, 2\|\Psi^{-1}E_n[\{1/2 - 1(U_i \leq 1/2)\}\widetilde{x}_i]\|_\infty), \tag{26}$$

where $Q(1 - \gamma, Z)$ denotes the $(1 - \gamma)$-quantile of a random variable Z. Here $U_1, \ldots, U_n$ are independent uniform random variables on $(0, 1)$ independent of $\widetilde{x}_1, \ldots, \widetilde{x}_n$. This quantity can be easily approximated via simulations. The values of γ and c_0 are chosen by users, but we suggest to take $\gamma = \gamma_n = 0{\cdot}1/\log n$ and $c_0 = 1{\cdot}1$. Below we summarize required technical conditions.

Condition 4. *Assume that $\|\eta_0\|_0 = s \geq 1$, $E(\widetilde{x}_{ij}^2) = 1$, $|E_n(\widetilde{x}_{ij}^2) - 1| \leq 1/2$ for $j = 1, \ldots, p$ with probability $1 - o(1)$, the conditional density of y_i given $\widetilde{x}_i$, denoted by $f_i(\cdot)$, and its derivative are bounded by $\bar{f}$ and $\bar{f}'$, respectively, and $f_i(\widetilde{x}_i^{\mathrm{T}}\eta_0) \geq \underline{f} > 0$ is bounded away from zero.*

Condition 4 is implied by Condition 1 after a normalizing the variables so that $E(\widetilde{x}_{ij}^2) = 1$ for $j = 1, \ldots, p$. The assumption on the conditional density is standard in the quantile regression literature even with fixed p or p increasing slower than n, see respectively [16] and [13].

We present bounds on the population prediction norm of the ℓ_1-penalized median regression estimator. The bounds depend on the restricted eigenvalue proposed in [7], defined by

$$\bar{\kappa}_{c_0} = \inf_{\delta \in \Delta_{c_0}} \|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}/\|\delta_{\widetilde{T}}\|,$$

where $\widetilde{T} = \mathrm{supp}(\eta_0)$, $\Delta_{c_0} = \{\delta \in \mathbb{R}^{p+1} : \|\delta_{\widetilde{T}^c}\|_1 \leq 3c_0'\|\delta_{\widetilde{T}}\|_1\}$ and $\widetilde{T}^c = \{1, \ldots, p+1\}\backslash\widetilde{T}$. The following lemma follows directly from the proof of Theorem 2 in [3] applied to a single quantile index.

Lemma 2. *Under Condition 4 and using* $\lambda = c_0 n\Lambda(1-\gamma \mid \widetilde{x}) \lesssim [n\log\{(p\vee n)/\gamma\}]^{1/2}$*, we have with probability at least* $1-\gamma-o(1)$*,*

$$\|\widetilde{x}_i^{\mathrm{T}}(\widehat{\eta}-\eta_0)\|_{P,2} \lesssim \frac{1}{\bar{\kappa}_{c_0}}\left[\frac{s\log\{(p\vee n)/\gamma\}}{n}\right]^{1/2},$$

provided that

$$\frac{n^{1/2}\bar{\kappa}_{c_0}}{[s\log\{(p\vee n)/\gamma\}]^{1/2}}\frac{\bar{f}\bar{f}'}{\underline{f}}\inf_{\delta\in\Delta_{c_0}}\frac{\|x^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} \to \infty.$$

Lemma 2 establishes the rate of convergence in the population prediction norm for the ℓ_1-penalized median regression estimator in a parametric setting. The extra growth condition required for identification is mild. For instance for many designs of interest we have

$$\inf_{\delta\in\Delta_{c_0}}\|x^{\mathrm{T}}\delta\|_{P,2}^3/E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)$$

bounded away from zero as shown in [3]. For designs with bounded regressors we have

$$\inf_{\delta\in\Delta_{c_0}}\frac{\|x^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} \geq \inf_{\delta\in\Delta_{c_0}}\frac{\|x^{\mathrm{T}}\delta\|_{P,2}}{\|\delta\|_1\widetilde{K}_n} \geq \frac{\bar{\kappa}_{c_0}}{s^{1/2}(1+3c_0')\widetilde{K}_n},$$

where $\widetilde{K}_n$ is a constant such that $\widetilde{K}_n \geq \|\widetilde{x}_i\|_\infty$ almost surely. This leads to the extra growth condition that $\widetilde{K}_n^2 s^2\log(p\vee n) = o(\bar{\kappa}_{c_0}^2 n)$.

In order to alleviate the bias introduced by the ℓ_1-penalty, we can consider the associated post-model selection estimate associated with a selected support $\widehat{T}$

$$\widetilde{\eta} \in \arg\min_{\eta}\left\{E_n(|y_i-\widetilde{x}_i^{\mathrm{T}}\eta|) : \mathrm{supp}(\eta)\subset\widehat{T}\right\}. \tag{27}$$

The following result characterizes the performance of the estimator in (27); see Theorem 5 in [3] for the proof.

Lemma 3. *Suppose that* $\mathrm{supp}(\widehat{\eta})\subset\widehat{T}$ *and let* $\widehat{s}=|\widehat{T}|$*. Then under the same conditions of Lemma 2,*

$$\|\widetilde{x}_i^{\mathrm{T}}(\widetilde{\eta}-\eta_0)\|_{P,2} \lesssim_P \left\{\frac{(\widehat{s}+s)\phi_{\max}(\widehat{s}+s)\log(n\vee p)}{n\bar{\phi}_{\min}(\widehat{s}+s)}\right\}^{1/2} + \frac{1}{\bar{\kappa}_{c_0}}\left[\frac{s\log\{(p\vee n)/\gamma\}}{n}\right]^{1/2},$$

provided that

$$n^{1/2}\frac{\{\bar{\phi}_{\min}(\widehat{s}+s)/\phi_{\max}(\widehat{s}+s)\}^{1/2}\wedge\bar{\kappa}_{c_0}}{[s\log\{(p\vee n)/\gamma\}]^{1/2}}\frac{\bar{f}\bar{f}'}{\underline{f}}\inf_{\|\delta\|_0\leq\widehat{s}+s}\frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} \to_P \infty.$$

Lemma 3 provides the rate of convergence in the prediction norm for the post model selection estimator despite possible imperfect model selection. The rates rely on the overall quality of the selected model, which is at least as good as the model selected by ℓ_1-penalized median regression, and the overall number of components $\widehat{s}$. Once again the extra growth condition required for identification is mild.

Comment D.1. In Step (i) of Algorithm 2 we use ℓ_1-penalized median regression with $\widetilde{x}_i = (d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\delta} = \widehat{\eta}-\eta_0 = (\widehat{\alpha}-\alpha_0, \widehat{\beta}^{\mathrm{T}}-\beta_0^{\mathrm{T}})^{\mathrm{T}}$, and we are interested in rates for $\|x_i^{\mathrm{T}}(\widehat{\beta}-\beta_0)\|_{P,2}$ instead of $\|\widetilde{x}_i^{\mathrm{T}}\widehat{\delta}\|_{P,2}$. However, it follows that

$$\|x_i^{\mathrm{T}}(\widehat{\beta}-\beta_0)\|_{P,2} \leq \|\widetilde{x}_i^{\mathrm{T}}\widehat{\delta}\|_{P,2} + |\widehat{\alpha}-\alpha_0|\,\|d_i\|_{P,2}.$$

Since $s \geq 1$, without loss of generality we can assume the component associated with the treatment d_i belongs to $\widetilde{T}$, at the cost of increasing the cardinality of $\widetilde{T}$ by one which will not affect the rate of convergence. Therefore we have that

$$|\widehat{\alpha} - \alpha_0| \leq \|\widehat{\delta}_{\widetilde{T}}\| \leq \|\widetilde{x}_i^{\mathrm{T}}\widehat{\delta}\|_{P,2}/\bar{\kappa}_{c0},$$

provided that $\widehat{\delta} \in \Delta_{c_0}$, which occurs with probability at least $1-\gamma$. In most applications of interest $\|d_i\|_{P,2}$ and $1/\bar{\kappa}_{c_0}$ are bounded from above. Similarly, in Step (i) of Algorithm 1 we have that the post-ℓ_1-penalized median regression estimator satisfies

$$\|x_i^{\mathrm{T}}(\widetilde{\beta} - \beta_0)\|_{P,2} \leq \|\widetilde{x}_i^{\mathrm{T}}\widetilde{\delta}\|_{P,2}\left[1 + \|d_i\|_{P,2}/\{\bar{\phi}_{\min}(\widehat{s}+s)\}^{1/2}\right].$$

D.3. **Heteroscedastic lasso.** In this section we consider the equation (4) of the form

$$d_i = x_i^{\mathrm{T}}\theta_0 + v_i,\ E(v_i \mid x_i) = 0 \quad (i = 1, \ldots, n),$$

where we observe $\{(d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ that are independent and identically distributed random vectors. The unknown support of θ_0 is denoted by T_d and it satisfies $|T_d| \leq s$. To estimate θ_0, we compute

$$\widehat{\theta} \in \arg\min_{\theta} E_n\{(d_i - x_i^{\mathrm{T}}\theta)^2\} + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1, \tag{28}$$

where λ and $\widehat{\Gamma}$ are the associated penalty level and loadings which are potentially data-driven. We rely on the results of [2] on the performance of lasso and post-lasso estimators that allow for heteroscedasticity and non-Gaussianity. According to [2], we use an initial and a refined option for the penalty level and the loadings, respectively

$$\begin{aligned} &\widehat{\gamma}_j = [E_n\{x_{ij}^2(d_i - \bar{d})^2\}]^{1/2}, \quad \lambda = 2cn^{1/2}\Phi^{-1}\{1-\gamma/(2p)\}, \\ &\widehat{\gamma}_j = \{E_n(x_{ij}^2\widehat{v}_i^2)\}^{1/2}, \qquad\quad \lambda = 2cn^{1/2}\Phi^{-1}\{1-\gamma/(2p)\}, \end{aligned} \tag{29}$$

for $j = 1, \ldots, p$, where $c > 1$ is a fixed constant, $\gamma \in (1/n, 1/\log n)$, $\bar{d} = E_n(d_i)$ and $\widehat{v}_i$ is an estimate of v_i based on lasso with the initial option or iterations.

We make the following high-level conditions. Below c_1, C_1 are given positive constants, and $\ell_n \uparrow \infty$ is a given sequence of constants.

Condition 5. *Suppose that (i) there exists $s = s_n \geq 1$ such that $\|\theta_0\|_0 \leq s$. (ii) $E(d^2) \leq C_1, \min_{j=1,\ldots,p} E(x_{ij}^2) \geq c_1$, $E(v^2 \mid x) \geq c_1$ almost surely, and $\max_{j=1,\ldots,p} E(|x_{ij}d_i|^2) \leq C_1$. (iii) $\max_{j=1,\ldots,p}\{E(|x_{ij}v_i|^3)\}^{1/3}\log^{1/2}(n \vee p) = o(n^{1/6})$. (iv) With probability $1 - o(1)$, $\max_{j=1,\ldots,p}|E_n(x_{ij}^2v_i^2) - E(x_{ij}^2v_i^2)| \vee \max_{j=1,\ldots,p}|E_n(x_{ij}^2d_i^2) - E(x_{ij}^2d_i^2)| = o(1)$ and $\max_{i=1,\ldots,n}\|x_i\|_\infty^2 s\log(n \vee p) = o(n)$. (v) With probability $1 - o(1)$, $c_1 \leq \phi_{\min}^x(\ell_n s) \leq \phi_{\max}^x(\ell_n s) \leq C_1$.*

Condition 5 (i) implies Condition AS in [2], while Conditions 5 (ii)-(iv) imply Condition RF in [2]. Lemma 3 in [2] provides primitive sufficient conditions under which condition (iv) is satisfied. The condition on the sparse eigenvalues ensures that $\kappa_{\bar{C}}$ in Theorem 1 of [2], applied to this setting, is bounded away from zero with probability $1 - o(1)$; see Lemma 4.1 in [7].

Next we summarize results on the performance of the estimators generated by lasso.

Lemma 4. *Suppose that Condition 5 is satisfied. Setting* $\lambda = 2cn^{1/2}\Phi^{-1}\{1-\gamma/(2p)\}$ *for* $c > 1$, *and using the penalty loadings as in (29), we have with probability* $1 - o(1)$,

$$\|x_i^{\mathrm{T}}(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim \frac{\lambda s^{1/2}}{n}.$$

Associated with lasso we can define the post-lasso estimator as

$$\widetilde{\theta} \in \arg\min_{\theta} \left\{ E_n\{(d_i - x_i^{\mathrm{T}}\theta)^2\} : \mathrm{supp}(\theta) \subset \mathrm{supp}(\widehat{\theta}) \right\}.$$

That is, the post-lasso estimator is simply the least squares estimator applied to the regressors selected by lasso in (28). Sparsity properties of the lasso estimator $\widehat{\theta}$ under estimated weights follows similarly to the standard lasso analysis derived in [2]. By combining such sparsity properties and the rates in the prediction norm, we can establish rates for the post-model selection estimator under estimated weights. The following result summarizes the properties of the post-lasso estimator.

Lemma 5. *Suppose that Condition 5 is satisfied. Consider the lasso estimator with penalty level and loadings specified as in Lemma 4. Then the data-dependent model* $\widehat{T}_d$ *selected by the lasso estimator* $\widehat{\theta}$ *satisfies with probability* $1 - o(1)$*:*

$$\|\widetilde{\theta}\|_0 = |\widehat{T}_d| \lesssim s.$$

Moreover, the post-lasso estimator obeys

$$\|x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \left\{ \frac{s \log(p \vee n)}{n} \right\}^{1/2}.$$

APPENDIX E. PROOFS FOR SECTION 2

E.1. **Proof of Theorem 1.** The proof of Theorem 1 consists of verifying Conditions 2 and 3 and application of Theorem 2. We will use the properties of the post-ℓ_1-penalized median regression and the post-lasso estimator together with required regularity conditions stated in Section D of this Supplementary Material. Moreover, we will use Lemmas 6 and 8 stated in Section G of this Supplementary Material. In this proof we focus on Algorithm 1. The proof for Algorithm 2 is essentially the same as that for Algorithm 1 and deferred to the next subsection.

In application of Theorem 2, take $p_1 = 1, z = x, w = (y, d, x^{\mathrm{T}})^{\mathrm{T}}, M = 2, \psi(w, \alpha, t) = \{1/2 - 1(y \leq \alpha d + t_1)\}(d - t_2), h(z) = (x^{\mathrm{T}}\beta_0, x^{\mathrm{T}}\theta_0)^{\mathrm{T}} = \{g(x), m(x)\}^{\mathrm{T}} = h(x)$, $\mathcal{A} = [\alpha_0 - c_2, \alpha_0 + c_2]$ where c_2 will be specified later, and $\mathcal{T} = \mathbb{R}^2$, we omit the subindex "$j$." In what follows, we will separately verify Conditions 2 and 3.

Verification of Condition 2: Part (i). The first condition follows from the zero median condition, that is, $F_\epsilon(0) = 1/2$. We will show in verification of Condition 3 that with probability $1 - o(1)$, $|\widehat{\alpha} - \alpha_0| = o(1/\log n)$, so that for some sufficiently small $c > 0$, $[\alpha_0 \pm c/\log n] \subset \widehat{\mathcal{A}} \subset \mathcal{A}$, with probability $1 - o(1)$.

Part (ii). The map

$$(\alpha, t) \mapsto E\{\psi(w, \alpha, t) \mid x\} = E([1/2 - F_\epsilon\{(\alpha - \alpha_0)d + t_1 - g(x)\}](d - t_2) \mid x)$$

is twice continuously differentiable since f'_ϵ is continuous. For every $\nu \in \{\alpha, t_1, t_2\}$, $\partial_\nu E\{\psi(w, \alpha, t) \mid x\}$ is $-E[f_\epsilon\{(\alpha - \alpha_0)d + t_1 - g(x)\}d(d - t_2) \mid x]$ or $-E[f_\epsilon\{(\alpha - \alpha_0)d + t_1 - g(x)\}(d - t_2) \mid x]$

or $E[F_\epsilon\{(\alpha-\alpha_0)d+t_1-g(x)\}\mid x]$. Hence for every $\alpha\in\mathcal{A}$,

$$|\partial_\nu E[\psi\{w,\alpha,h(x)\}\mid x]|\le C_1E(|dv|\mid x)\vee C_1E(|v|\mid x)\vee 1.$$

The expectation of the square of the right side is bounded by a constant depending only on c_3,C_1, as $E(d^4)+E(v^4)\le C_1$. Moreover, let $\mathcal{T}(x)=\{t\in\mathbb{R}^2:|t_2-m(x)|\le c_3\}$ with any fixed constant $c_3>0$. Then for every $\nu,\nu'\in\{\alpha,t,t'\}$, whenever $\alpha\in\mathcal{A},t\in\mathcal{T}(x)$,

$$\begin{aligned}&|\partial_\nu\partial_{\nu'}E\{\psi(w,\alpha,t)\mid x\}|\\&\le C_1\left[1\vee E\{|d^2(d-t_2)|\mid x\}\vee E\{|d(d-t_2)|\mid x\}\vee E(|d|\mid x)\vee E(|d-t_2|\mid x)\right].\end{aligned}$$

Since $d=m(x)+v,|m(x)|=|x^{\mathrm{T}}\theta_0|\le M_n,|t_2-m(x)|\le c_3$ for $t\in\mathcal{T}(x)$, and $E(|v|^3\mid x)\le C_1$, we have

$$\begin{aligned}&E\{|d^2(d-t_2)|\mid x\}\le E[\{m(x)+v\}^2(c_3+|v|)\mid x]\le 2E[\{m^2(x)+v^2\}(c_3+|v|)\mid x]\\&\le 2E\{(M_n^2+v^2)(c_3+|v|)\mid x\}\lesssim M_n^2.\end{aligned}$$

Similar computations lead to $|\partial_\nu\partial_{\nu'}E\{\psi(w,\alpha,t)\mid x\}|\le CM_n^2=L_{1n}$ for some constant C depending only on c_3,C_1. We wish to verify the last condition in (ii). For every $\alpha,\alpha'\in\mathcal{A},t,t'\in\mathcal{T}(x)$,

$$\begin{aligned}&E[\{\psi(w,\alpha,t)-\psi(w,\alpha',t')\}^2\mid x]\le C_1E\{|d(d-t_2)|\mid x\}|\alpha-\alpha'|\\&\quad+C_1E\{|(d-t_2)|\mid x\}|t_1-t_1'|+(t_2-t_2')^2\le C'M_n(|\alpha-\alpha'|+|t_1-t_1'|)+(t_2-t_2')^2,\end{aligned}$$

where C' is a constant depending only on c_3,C_1. Here as $|t_2-t_2'|\le|t_2-m(x)|+|m(x)-t_2|\le 2c_3$, the right side is bounded by $2^{1/2}(C'M_n+2c_3)(|\alpha-\alpha'|+\|t-t'\|)$. Hence we can take $L_{2n}=2^{1/2}(C'M_n+2c_3)$ and $\varsigma=1$.

Part (iii). Recall that $d=x^{\mathrm{T}}\theta_0+v,E(v\mid x)=0$. Then we have

$$\begin{aligned}\partial_{t_1}E\{\psi(w,\alpha_0,t)\mid x\}|_{t=h(x)}&=E\{f_\epsilon(0)v\mid x\}=0,\\\partial_{t_2}E\{\psi(w,\alpha_0,t)\mid x\}|_{t=h(x)}&=-E\{F_\epsilon(0)-1/2\mid x\}=0.\end{aligned}$$

Part (iv). Pick any $\alpha\in\mathcal{A}$. There exists α' between α_0 and α such that

$$E[\psi\{w,\alpha,h(x)\}]=\partial_\alpha E[\psi\{w,\alpha_0,h(x)\}](\alpha-\alpha_0)+\frac{1}{2}\partial_\alpha^2E[\psi\{w,\alpha',h(x)\}](\alpha-\alpha_0)^2$$

Let $\Gamma=\partial_\alpha E[\psi\{w,\alpha_0,h(x)\}]=f_\epsilon(0)E(v^2)\ge c_1^2$. Then since $|\partial_\alpha^2E[\psi\{w,\alpha',h(x)\}]|\le C_1E(|d^2v|)\le C_2$ where C_2 can be taken depending only on C_1, we have

$$E[\psi\{w,\alpha,h(x)\}]\ge\frac{1}{2}\Gamma|\alpha-\alpha_0|,$$

whenever $|\alpha-\alpha_0|\le c_1^2/C_2$. Take $c_2=c_1^2/C_2$ in the definition of $\mathcal{A}$. Then the above inequality holds for all $\alpha\in\mathcal{A}$.

Part (v). Observe that $E[\psi^2\{w,\alpha_0,h(x)\}]=(1/4)E(v^2)\ge c_1/4$.

Verification of Condition 3: Note here that $a_n=p\vee n$ and $b_n=1$. We first show that the estimators $\widehat{h}(x)=(x^{\mathrm{T}}\widetilde{\beta},x^{\mathrm{T}}\widetilde{\theta})^{\mathrm{T}}$ are sparse and have good rate properties.

The estimator $\widetilde{\beta}$ is based on post-ℓ_1-penalized median regression with penalty parameters as suggested in Section D.2 of this Supplementary Material. By assumption in Theorem 1, with probability $1-\Delta_n$ we have $\widehat{s}=\|\widetilde{\beta}\|_0\le C_1s$. Next we verify that Condition 4 in Section D.2 of this Supplementary Material is implied by Condition 1 and invoke Lemmas 2 and 3.

The assumptions on the error density $f_\epsilon(\cdot)$ in Condition 4 (i) are assumed in Condition 1 (iv). Because of Conditions 1 (v) and (vi), $\bar{\kappa}_{c_0}$ is bounded away from zero for n sufficiently large, see Lemma 4.1 in [7], and $c_1 \leq \bar{\phi}_{\min}(1) \leq E(\widetilde{x}_j^2) \leq \bar{\phi}_{\max}(1) \leq C_1$ for every $j = 1, \ldots, p$. Moreover, under Condition 1, by Lemma 8, we have $\max_{j=1,\ldots,p+1} |E_n(\widetilde{x}_j^2)/E(\widetilde{x}_j^2) - 1| \leq 1/2$ and $\phi_{\max}(\ell_n' s) \leq 2E_n(d^2) + 2\phi_{\max}^x(\ell_n' s) \leq 5C_1$ with probability $1 - o(1)$ for some $\ell_n' \to \infty$. The required side condition of Lemma 2 is satisfied by relations (30) and (31) ahead. By Lemma 3 in Section D.2 of this Supplementary Material, we have $\|x_i^{\mathrm{T}}(\widetilde{\beta} - \beta_0)\|_{P,2} \lesssim_P \{s \log(n \vee p)/n\}^{1/2}$ since the required side condition holds. Indeed, for $\widetilde{x}_i = (d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$ and $\delta = (\delta_d, \delta_x^{\mathrm{T}})^{\mathrm{T}}$, because $\|\widetilde{\beta}\|_0 \leq C_1 s$ with probability $1 - \Delta_n$, $c_1 \leq \bar{\phi}_{\min}(C_1 s + s) \leq \bar{\phi}_{\max}(C_1 s + s) \leq C_1$, and $E(|d_i|^3) = O(1)$, we have

$$\begin{aligned}
\inf_{\|\delta\|_0 \leq s + C_1 s} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} &\geq \inf_{\|\delta\|_0 \leq s + C_1 s} \frac{\{\bar{\phi}_{\min}(s + C_1 s)\}^{3/2}\|\delta\|^3}{4E(|x_i^{\mathrm{T}}\delta_x|^3) + 4|\delta_d|^3 E(|d_i|^3)} \\
&\geq \inf_{\|\delta\|_0 \leq s + C_1 s} \frac{\{\bar{\phi}_{\min}(s + C_1 s)\}^{3/2}\|\delta\|^3}{4K_n\|\delta_x\|_1 \bar{\phi}_{\max}(s + C_1 s)\|\delta_x\|^2 + 4\|\delta\|^3 E(|d_i|^3)} \\
&\geq \frac{\{\bar{\phi}_{\min}(s + C_1 s)\}^{3/2}}{4K_n\{s + C_1 s\}^{1/2}\phi_{\max}(s + C_1 s) + 4E(|d_i|^3)} \gtrsim \frac{1}{K_n s^{1/2}}.
\end{aligned}$$

Therefore, since $K_n^2 s^2 \log^2(p \vee n) = o(n)$, we have

$$n^{1/2} \frac{\{\bar{\phi}_{\min}(s + C_1 s)/\phi_{\max}(s + C_1 s)\}^{1/2} \wedge \bar{\kappa}_{c_0}}{\{s \log(p \vee n)\}^{1/2}} \inf_{\|\delta\|_0 \leq s + C_1 s} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} \gtrsim \frac{n^{1/2}}{K_n s \log(p \vee n)} \to \infty.$$

The argument above also shows that $|\widehat{\alpha} - \alpha_0| = o(1/\log n)$ with probability $1 - o(1)$ as claimed in Verification of Condition 2 (i). Indeed by Lemma 2 and Remark D.1 we have $|\widehat{\alpha} - \alpha_0| \lesssim \{s \log(p \vee n)/n\}^{1/2} = o(1/\log n)$ with probability $1 - o(1)$ as $s^2 \log^3(p \vee n) = o(n)$.

The $\widetilde{\theta}$ is a post-lasso estimator with penalty parameters as suggested in Section D.3 of this Supplementary Material. We verify that Condition 5 in Section D.3 of this Supplementary Material is implied by Condition 1 and invoke Lemma 5. Indeed, Condition 5 (ii) is implied by Conditions 1 (ii) and (iv), where Condition 1(iv) is used to ensure $\min_{j=1,\ldots,p} E(x_j^2) \geq c_1$. Next since $\max_{j=1,\ldots,p} E(|x_j v|^3) \leq C_1$, Condition 5 (iii) is satisfied if $\log^{1/2}(p \vee n) = o(n^{1/6})$, which is implied by Condition 1 (v). Condition 5 (iv) follows from Lemma 6 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ as $K_n^4 \log p = o(n)$ and $K_n^2 s \log(p \vee n) = o(n)$. Condition 5 (v) follows from Lemma 8. By Lemma 5 in Section D.3 of this Supplementary Material, we have $\|x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \{s \log(n \vee p)/n\}^{1/2}$ and $\|\widetilde{\theta}\|_0 \lesssim s$ with probability $1 - o(1)$. Thus, by Lemma 8, we have $\|x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{P,2} \lesssim_P \{s \log(n \vee p)/n\}^{1/2}$. Moreover, $\sup_{\|x\|_\infty \leq K_n} |x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)| \leq K_n\|\widetilde{\theta} - \theta_0\|_1 \leq K_n s^{1/2}\|\widetilde{\theta} - \theta_0\| \lesssim_P K_n s\{\log(n \vee p)/n\}^{1/2} = o(1)$.

Combining these results, we have $\widehat{h} \in \mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ with probability $1 - o(1)$, where

$$\begin{aligned}
\mathcal{H}_1 &= \{\widetilde{h}_1 : \widetilde{h}_1(x) = x^{\mathrm{T}}\beta, \|\beta\|_0 \leq C_3 s, E[\{\widetilde{h}_1(x) - g(x)\}^2] \leq \ell_n' s(\log a_n)/n\}, \\
\mathcal{H}_2 &= \{\widetilde{h}_2 : \widetilde{h}_2(x) = x^{\mathrm{T}}\theta, \|\theta\|_0 \leq C_3 s, \sup_{\|x\|_\infty \leq K_n} |\widetilde{h}_2(x) - m(x)| \leq c_3, \\
&\qquad E[\{\widetilde{h}_2(x) - m(x)\}^2] \leq \ell_n' s(\log a_n)/n\},
\end{aligned}$$

with C_3 a sufficiently large constant and $\ell_n' \uparrow \infty$ sufficiently slowly.

To verify Condition 3 (ii), observe that $\mathcal{F} = \varphi(\mathcal{G}) \cdot \mathcal{G}'$, where $\varphi(u) = 1/2 - 1(u \leq 0)$, and $\mathcal{G}$ and $\mathcal{G}'$ are the classes of functions defined by

$$\mathcal{G} = \{(y, d, x^{\mathrm{T}})^{\mathrm{T}} \mapsto y - \alpha d - \widetilde{h}_1(x) : \alpha \in \mathcal{A}, \widetilde{h}_1 \in \mathcal{H}_1\},$$
$$\mathcal{G}' = \{(y, d, x^{\mathrm{T}})^{\mathrm{T}} \mapsto d - \widetilde{h}_2(x) : \widetilde{h}_2 \in \mathcal{H}_2\}.$$

The classes $\mathcal{G}, \mathcal{G}'$, and $\varphi(\mathcal{G})$, as φ is monotone and by Lemma 2.6.18 in [32], consist of unions of p choose $C_3 s$ VC-subgraph classes with VC indices at most $C_3 s + 3$. The class $\varphi(\mathcal{G})$ is uniformly bounded by 1; recalling $d = m(x) + v$, for $\widetilde{h}_2 \in \mathcal{H}_2$, $|d - \widetilde{h}_2(x)| \leq c_3 + |v|$. Hence by Theorem 2.6.7 in [32], we have $\mathrm{ent}\{\varepsilon, \varphi(\mathcal{G})\} \vee \mathrm{ent}(\varepsilon, \mathcal{G}') \leq C'' s \log(a_n/\varepsilon)$ for all $0 < \varepsilon \leq 1$ for some constant C'' that depends only on C_3; see the proof of Lemma 11 in [3] for related arguments. It is now straightforward to verify that the class $\mathcal{F} = \varphi(\mathcal{G}) \cdot \mathcal{G}'$ satisfies the stated entropy condition; see the proof of Theorem 3 in [1], relation (A.7).

To verify Condition 3 (iii), observe that whenever $\widetilde{h}_2 \in \mathcal{H}_2$,

$$|\varphi\{y - \alpha d - \widetilde{h}_1(x)\}\{d - \widetilde{h}_2(x)\}| \leq c_3 + |v|,$$

which has four bounded moments, so that Condition 3 (iii) is satisfied with $q = 4$.

To verify Condition 3 (iv), take $s = \ell_n' s$ with $\ell_n' \uparrow \infty$ sufficiently slowly and

$$\rho_n = n^{-1/2}\{(\ell_n' s \log a_n)^{1/2} + n^{-1/4}\ell_n' s \log a_n\} \lesssim n^{-1/2}(\ell_n' s \log a_n)^{1/2}.$$

As $\varsigma = 1$, $L_{1n} \lesssim M_n^2$ and $L_{2n} \lesssim M_n$, Condition 3 (iv) is satisfied provided that $M_n^2 s_n^3 \log^3 a_n = o(n)$ and $M_n^4 s_n^2 \log^2 a_n = o(n)$, which are implied by Condition 1 (v) with $\ell_n' \uparrow \infty$ sufficiently slowly.

Therefore, for $\sigma_n^2 = E[\Gamma^{-2}\psi\{w, \alpha_0, h(x)\}] = E(v_i^2)/\{4f_\epsilon^2(0)\}$, by Theorem 2 we obtain the first result: $\sigma_n^{-1} n^{1/2}(\check{\alpha} - \alpha_0) \to \mathcal{N}(0, 1)$.

Next we prove the second result regarding $nL_n(\alpha_0)$. First consider the denominator of $L_n(\alpha_0)$. We have

$$\begin{aligned}|E_n(\widehat{v}_i^2) - E_n(v_i^2)| &= |E_n\{(\widehat{v}_i - v_i)(\widehat{v}_i + v_i)\}| \leq \|\widehat{v}_i - v_i\|_{2,n}\|\widehat{v}_i + v_i\|_{2,n} \\ &\leq \|x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n}\{2\|v_i\|_{2,n} + \|x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n}\} = o_P(1),\end{aligned}$$

where we have used $\|v_i\|_{2,n} \lesssim_P \{E(v_i^2)\}^{1/2} = O(1)$ and $\|x_i^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n} = o_P(1)$.

Second consider the numerator of $L_n(\alpha_0)$. Since $E[\psi\{w, \alpha_0, h(x)\}] = 0$ we have

$$E_n[\psi\{w, \alpha_0, \widehat{h}(x)\}] \quad = E_n[\psi\{w, \alpha_0, h(x)\}] + o_P(n^{-1/2}),$$

using the expansion in the displayed equation of *Step* 1 in the proof of Theorem 2 evaluated at α_0 instead of $\widetilde{\alpha}_j$. Therefore, using the identity that $nA_n^2 = nB_n^2 + n(A_n - B_n)^2 + 2nB_n(A_n - B_n)$ with

$$A_n = E_n[\psi\{w, \alpha_0, \widehat{h}(x)\}],\ B_n = E_n[\psi\{w, \alpha_0, h(x)\}],\ |B_n| \lesssim_P \{E(v_i^2)\}^{1/2} n^{-1/2},$$

we have

$$nL_n(\alpha_0) = \frac{4n|E_n[\psi\{w, \alpha_0, \widehat{h}(x)\}]|^2}{E_n(\widehat{v}_i^2)} = \frac{4n|E_n[\psi\{w, \alpha_0, h(x)\}]|^2}{E_n[\psi^2\{w, \alpha_0, h(x)\}]} + o_P(1)$$

since $E(v_i^2)$ is bounded away from zero. By Theorem 7.1 in [12], and the moment conditions $E(d^4) \leq C_1$ and $E(v^2) \geq c_1$, the following holds for the self-normalized sum

$$I = \frac{2n^{1/2}E_n[\psi\{w, \alpha_0, h(x)\}]}{(E_n[\psi^2\{w, \alpha_0, h(x)\}])^{1/2}} \to \mathcal{N}(0,1)$$

in distribution and the desired result follows since $nL_n(\alpha_0) = I^2 + o_P(1)$.

Comment E.1. An inspection of the proof leads to the following stochastic expansion:

$$\begin{aligned} E_n[\psi\{w, \widehat{\alpha}, \widehat{h}(x)\}] = &-\{f_\epsilon E(v_i^2)\}(\widehat{\alpha} - \alpha_0) + E_n[\psi\{w, \alpha_0, h(x)\}] \\ &+ o_P(n^{-1/2} + n^{-1/4}|\widehat{\alpha} - \alpha_0|) + O_P(|\widehat{\alpha} - \alpha_0|^2), \end{aligned}$$

where $\widehat{\alpha}$ is any consistent estimator of α_0. Hence provided that $|\widehat{\alpha} - \alpha_0| = o_P(n^{-1/4})$, the remainder term in the above expansion is $o_P(n^{-1/2})$, and the one-step estimator $\check{\alpha}$ defined by

$$\check{\alpha} = \widehat{\alpha} + \{E_n(f_\epsilon \widehat{v}_i^2)\}^{-1} E_n[\psi\{w, \widehat{\alpha}, \widehat{h}(x)\}]$$

has the following stochastic expansion:

$$\begin{aligned} \check{\alpha} &= \widehat{\alpha} + \{f_\epsilon E(v_i^2) + o_P(n^{-1/4})\}^{-1}[-\{f_\epsilon E(v_i^2)\}(\widehat{\alpha} - \alpha_0) + E_n[\psi\{w, \alpha_0, h(x)\}] + o_P(n^{-1/2})] \\ &= \alpha_0 + \{f_\epsilon E(v_i^2)\}^{-1} E_n[\psi\{w, \alpha_0, h(x)\}] + o_P(n^{-1/2}), \end{aligned}$$

so that $\sigma_n^{-1} n^{1/2}(\check{\alpha} - \alpha_0) \to \mathcal{N}(0,1)$ in distribution.

E.2. **Proof of Theorem 1: Algorithm 2.**

Proof of Theorem 1: Algorithm 2. The proof is essentially the same as the proof for Algorithm 1 and just verifying the rates for the penalized estimators.

The estimator $\widehat{\beta}$ is based on ℓ_1-penalized median regression. Condition 4 is implied by Condition 1, see the proof for Algorithm 1. By Lemma 2 and Remark D.1 we have with probability $1 - o(1)$

$$\|x_i^{\mathrm{T}}(\widehat{\beta} - \beta_0)\|_{P,2} \lesssim \{s \log(n \vee p)/n\}^{1/2}, \quad |\widehat{\alpha} - \alpha_0| \lesssim \{s \log(p \vee n)/n\}^{1/2} = o(1/\log n),$$

because $s^3 \log^3(n \vee p) = o(n)$ and the required side condition holds. Indeed, without loss of generality assume that $\widetilde{T}$ contains d so that for $\widetilde{x}_i = (d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$, $\delta = (\delta_d, \delta_x^{\mathrm{T}})^{\mathrm{T}}$, because $\bar{\kappa}_{c_0}$ is bounded away from zero, and the fact that $E(|d_i|^3) = O(1)$, we have

$$\begin{aligned} \inf_{\delta \in \Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} &\geq \inf_{\delta \in \Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^2 \|\delta_T\| \bar{\kappa}_{c_0}}{4E(|x_i'\delta_x|^3) + 4E(|d_i\delta_d|^3)} \\ &\geq \inf_{\delta \in \Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^2 \|\delta_T\| \bar{\kappa}_{c_0}}{4K_n \|\delta_x\|_1 E(|x_i^{\mathrm{T}}\delta_x|^2) + 4|\delta_d|^3 E(|d_i|^3)} \\ &\geq \inf_{\delta \in \Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^2 \|\delta_T\| \bar{\kappa}_{c_0}}{\{4K_n \|\delta_x\|_1 + 4|\delta_d| E(|d_i|^3)/E(|d_i|^2)\}\{E(|x_i^{\mathrm{T}}\delta_x|^2) + E(|\delta_d d_i|^2)\}} \\ &\geq \inf_{\delta \in \Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^2 \|\delta_T\| \bar{\kappa}_{c_0}}{8(1+3c_0')\|\delta_T\|_1\{K_n + O(1)\}\{2E(|\widetilde{x}_i^{\mathrm{T}}\delta_x|^2) + 3E(|\delta_d d_i|^2)\}} \\ &\geq \inf_{\delta \in \Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^2 \|\delta_T\| \bar{\kappa}_{c_0}}{8(1+3c_0')\|\delta_T\|_1\{K_n + O(1)\}E(|\widetilde{x}_i^{\mathrm{T}}\delta_x|^2)(2 + 3/\bar{\kappa}_{c_0}^2)} \\ &\geq \frac{\bar{\kappa}_{c_0}/s^{1/2}}{8\{K_n + O(1)\}(1+3c_0')\{2 + 3E(d^2)/\bar{\kappa}_{c_0}^2\}} \gtrsim \frac{1}{s^{1/2}K_n}. \end{aligned} \tag{30}$$

Therefore, since $K_n^2 s^2 \log^2(p \vee n) = o(n)$, we have

$$\frac{n^{1/2}\bar{\kappa}_{c_0}}{\{s\log(p\vee n)\}^{1/2}} \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}_i^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}_i^{\mathrm{T}}\delta|^3)} \gtrsim \frac{n^{1/2}}{K_n s\log^{1/2}(p\vee n)} \to \infty. \tag{31}$$

The estimator $\widehat{\theta}$ is based on lasso. Condition 5 is implied by Condition 1 and Lemma 6 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ as $K_n^4 \log p = o(n)$. By Lemma 4 we have $\|x_i^{\mathrm{T}}(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \{s\log(n\vee p)/n\}^{1/2}$. Moreover, by Lemma 5 we have $\|\widehat{\theta}\|_0 \lesssim s$ with probability $1 - o(1)$. The required rate in the $\|\cdot\|_{P,2}$ norm follows from Lemma 8.

□

APPENDIX F. ADDITIONAL MONTE-CARLO EXPERIMENTS

In this section we provide additional experiments to further examine the finite sample performance of the proposed estimators. The experiments investigate the performance of the method on approximately spare models and complement the experiments on exactly sparse models presented in the main text. Specifically, we considered the following regression model:

$$y = d\alpha_0 + x^{\mathrm{T}}(c_y\theta_0) + \epsilon, \quad d = x^{\mathrm{T}}(c_d\theta_0) + v, \tag{32}$$

where $\alpha_0 = 1/2$, and now we have $\theta_{0j} = 1/j^2, j = 1,\ldots,p$. The other features of the design are the same as the design presented in the main text. Namely, the vector $x = (1, z^{\mathrm{T}})^{\mathrm{T}}$ consists of an intercept and covariates $z \sim N(0,\Sigma)$, and the errors ϵ and v are independently and identically distributed as $\mathcal{N}(0,1)$. The dimension p of the covariates x is 300, and the sample size n is 250. The regressors are correlated with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0{\cdot}5$. We vary the R^2 in the two equations, denoted by R_y^2 and R_d^2 respectively, in the set $\{0, 0{\cdot}1, \ldots, 0{\cdot}9\}$, which results in 100 different designs induced by the different pairs of (R_y^2, R_d^2). We performed 500 Monte Carlo repetitions for each.

In this design, the vector θ_0 has all p components different from zero. Because the coefficients decay it is conceivable that it can be well approximated by considering only a few components, typically the ones associated with the largest coefficients in absolute values. The coefficients omitted from that construction define the approximation error. However, the number of coefficients needed to achieve a good approximation will also depend on the scalings c_y and c_d since they multiply all coefficients. Therefore, if c_y or c_d is large the approximation might require a larger number of coefficients which can violate our sparsity requirements. This is the main distinction from the an exact sparse designs considered in the main text.

The simulation study focuses on Algorithm 1 since the algorithm based on double selection worked similarly. Standard errors are computed using the formula (11). As the main benchmark we consider the standard post-model selection estimator $\widetilde{\alpha}$ based on the post-ℓ_1-penalized median regression method, as defined in (3).

Figure 3 displays the empirical rejection probability of tests of a true hypothesis $\alpha = \alpha_0$, with nominal size of tests equal to 5%. The rejection frequency of the standard post-model selection inference procedure based upon $\widetilde{\alpha}$ is very fragile, see left plot. Given the approximately sparse model considered here, there is no true model to be perfectly recovered and the rejection frequency deviates substantially from the ideal rejection frequency of 5%. The right plot shows the corresponding empirical rejection probability for the proposed procedures based on estimator $\check{\alpha}$ and the result (10). The performance is close to the ideal level of 5% over 99 out of the 100

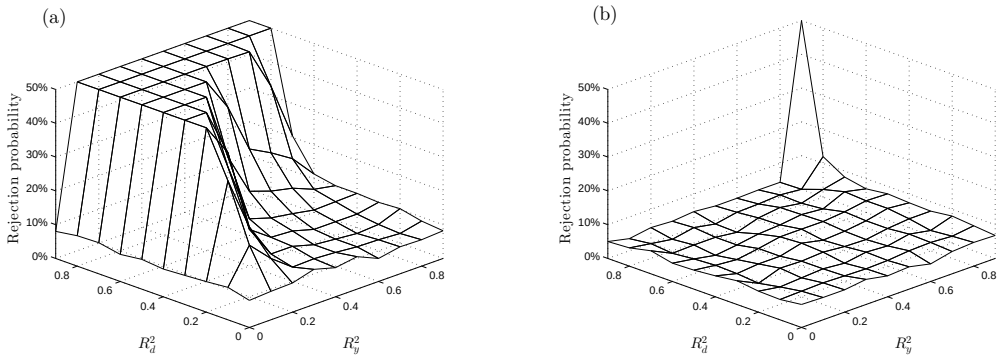

FIGURE 3. The empirical rejection probabilities of the nominal 5% level tests of a true hypothesis based on: (a) the standard post-model selection procedure based on $\widetilde{\alpha}$, and (b) the proposed post-model selection procedure based on $\check{\alpha}$. Ideally we should observe a flat surface at the 5% rejection rate (of a true null).

designs considered in the study which illustrate the uniformity property. The design for which the procedure does not perform well corresponds to $R_d^2 = 0{\cdot}9$ and $R_y^2 = 0{\cdot}9$.

Figure 4 compares the performance of the standard post-selection estimator $\widetilde{\alpha}$, as defined in (3), and our proposed post-selection estimator $\check{\alpha}$ obtained via Algorithm 1. We display results in the same three metrics used in the main text: mean bias, standard deviation, and root mean square error of the two approaches. In those metrics, except for one design, the performance for approximately sparse models is very similar to the performance of exactly sparse models. The proposed post-selection estimator $\check{\alpha}$ performs well in all three metrics while the standard post-model selection estimators $\widetilde{\alpha}$ exhibits a large bias in many of the dgps considered. For the design with $R_d^2 = 0{\cdot}9$ and $R_y^2 = 0{\cdot}9$, both procedures breakdown.

Except for the design with largest values of R^2's, $R_d^2 = 0{\cdot}9$ and $R_y^2 = 0{\cdot}9$, the results are very similar to the results presented in the main text for an exactly sparse model where the proposed procedure performs very well. The design with the largest values of R^2's correspond to large values of c_y and c_d. In that case too many coefficients are needed to achieve a good approximation for the unknown functions $x^{\mathrm{T}}(c_y\theta_0)$ and $x^{\mathrm{T}}(c_d\theta_0)$ which translates into a (too) large value of s in the approximate sparse model. Such performance is fully consistent with the theoretical result derived in Theorem 2 which covers approximately sparse models but do impose sparsity requirements.

APPENDIX G. AUXILIARY TECHNICAL RESULTS

In this section we collect some auxiliary technical results.

Lemma 6. *Let $(\zeta_1, x_1^{\mathrm{T}})^{\mathrm{T}}, \ldots, (\zeta_n, x_n^{\mathrm{T}})^{\mathrm{T}}$ be independent random vectors where $\zeta_1, \ldots, \zeta_n$ are scalar while $x_1, \ldots, x_n$ are vectors in $\mathbb{R}^p$. Suppose that $E(\zeta_i^4) < \infty$ for $i = 1, \ldots, n$, and there exists a constant K_n such that $\max_{i=1,\ldots,n} \|x_i\|_\infty \leq K_n$ almost surely. Then for every $\tau \in (0, 1/8)$, with probability at least $1 - 8\tau$,*

$$\max_{j=1,\ldots,p} |n^{-1}\textstyle\sum_{i=1}^n \{\zeta_i^2 x_{ij}^2 - E(\zeta_i^2 x_{ij}^2)\}| \leq 4K_n^2\{(2/n)\log(2p/\tau)\}^{1/2}\{\textstyle\sum_{i=1}^n E(\zeta_i^4)/(n\tau)\}^{1/2}.$$

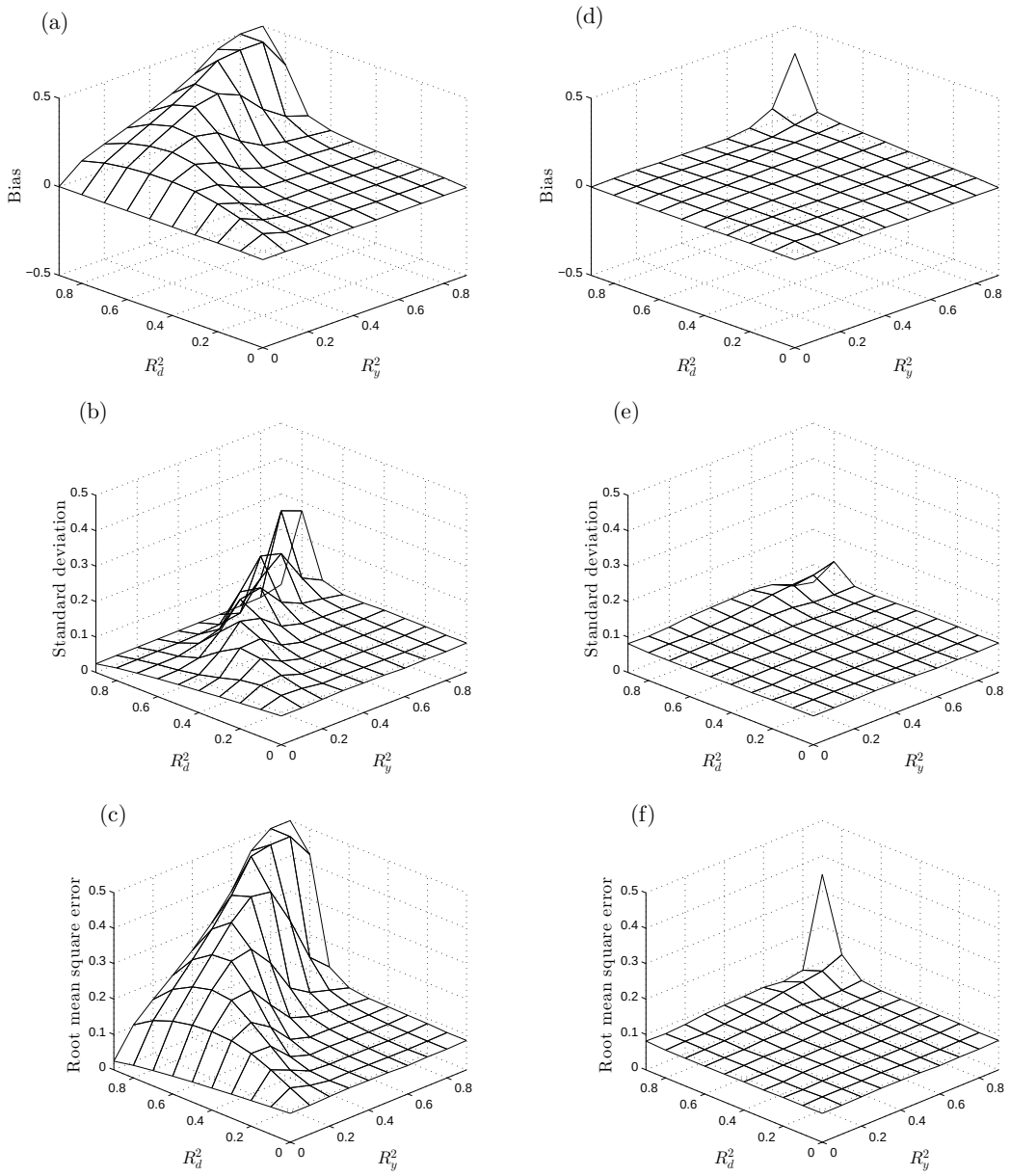

FIGURE 4. Mean bias (top row), standard deviation (middle row), root mean square error (bottom row) of the standard post-model selection estimator $\widetilde{\alpha}$ (panels (a)-(c)), and of the proposed post-model selection estimator $\check{\alpha}$ (panels (d)-(f)).

Proof of Lemma 6. The proof depends on the following maximal inequality derived in [5].

Lemma 7. *Let $z_1, \ldots, z_n$ be independent random vectors in $\mathbb{R}^p$. Then for every $\tau \in (0, 1/4)$ and $\delta \in (0, 1/4)$, with probability at least $1 - 4\tau - 4\delta$,*

$$\max_{j=1,\ldots,p} |n^{-1/2}\textstyle\sum_{i=1}^n \{z_{ij} - E(z_{ij})\}| \leq \left[4\{2\log(2p/\delta)\}^{1/2}\, Q\{1-\tau, \max_{j=1,\ldots,p}(n^{-1}\textstyle\sum_{i=1}^n z_{ij}^2)^{1/2}\}\right] \vee 2 \max_{j=1,\ldots,p} Q[1/2, |n^{-1/2}\textstyle\sum_{i=1}^n \{z_{ij} - E(z_{ij})\}|],$$

where for a random variable Z we denote $Q(u, Z) = u$-quantile of Z .

Going back to the proof of Lemma 6, let $z_{ij} = \zeta_i^2 x_{ij}^2$. By Markov's inequality, we have

$$Q[1/2, |n^{-1/2}\textstyle\sum_{i=1}^n \{z_{ij} - E(z_{ij})\}|] \leq \{2n^{-1}\textstyle\sum_{i=1}^n E(z_{ij}^2)\}^{1/2} \leq K_n^2\{(2/n)\textstyle\sum_{i=1}^n E(\zeta_i^4)\}^{1/2},$$

and

$$\begin{aligned} Q\{1-\tau, \max_{j=1,\ldots,p}(n^{-1}\textstyle\sum_{i=1}^n z_{ij}^2)^{1/2}\} &\leq Q\{1-\tau, K_n^2(n^{-1}\textstyle\sum_{i=1}^n \zeta_i^4)^{1/2}\} \\ &\leq K_n^2\{\textstyle\sum_{i=1}^n E(\zeta_i^4)/(n\tau)\}^{1/2}. \end{aligned}$$

Hence the conclusion of Lemma 6 follows from application of Lemma 7 with $\tau = \delta$. □

Lemma 8. *Under Condition 1, there exists $\ell_n' \to \infty$ such that with probability $1 - o(1)$,*

$$\sup_{\substack{\|\delta\|_0 \leq \ell_n' s \\ \delta \neq 0}} \left| \frac{\|x_i^{\mathrm{T}}\delta\|_{2,n}}{\|x_i^{\mathrm{T}}\delta\|_{P,2}} - 1 \right| = o(1).$$

Proof of Lemma 8. The lemma follows from application of Theorem 4.3 in [29].

□

Lemma 9. *Consider vectors $\widehat{\beta}$ and β_0 in $\mathbb{R}^p$ where $\|\beta_0\|_0 \leq s$, and denote by $\widehat{\beta}^{(m)}$ the vector $\widehat{\beta}$ truncated to have only its $m \geq s$ largest components in absolute value. Then*

$$\begin{aligned} &\|\widehat{\beta}^{(m)} - \beta_0\|_1 \leq 2\|\widehat{\beta} - \beta_0\|_1 \\ &\|x_i^{\mathrm{T}}\{\widehat{\beta}^{(2m)} - \beta_0\}\|_{2,n} \leq \|x_i^{\mathrm{T}}(\widehat{\beta} - \beta_0)\|_{2,n} + \{\phi_{\max}^x(m)/m\}^{1/2}\|\widehat{\beta} - \beta_0\|_1. \end{aligned}$$

Proof of Lemma 9. The first inequality follows from the triangle inequality

$$\|\widehat{\beta}^{(m)} - \beta_0\|_1 \leq \|\widehat{\beta} - \widehat{\beta}^{(m)}\|_1 + \|\widehat{\beta} - \beta_0\|_1$$

and the observation that $\|\widehat{\beta} - \widehat{\beta}^{(m)}\|_1 = \min_{\|\beta\|_0 \leq m} \|\widehat{\beta} - \beta\|_1 \leq \|\widehat{\beta} - \beta_0\|_1$ since $m \geq s = \|\beta_0\|_0$.

By the triangle inequality we have

$$\|x_i^{\mathrm{T}}\{\widehat{\beta}^{(2m)} - \beta_0\}\|_{2,n} \leq \|x_i^{\mathrm{T}}(\widehat{\beta} - \beta_0)\|_{2,n} + \|x_i^{\mathrm{T}}\{\widehat{\beta}^{(2m)} - \widehat{\beta}\}\|_{2,n}.$$

For an integer $k \geq 2$, $\|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_0 \leq m$ and $\widehat{\beta} - \widehat{\beta}^{(2m)} = \sum_{k \geq 3}\{\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\}$. Moreover, given the monotonicity of the components,

$$\|\widehat{\beta}^{(km+m)} - \widehat{\beta}^{(km)}\| \leq \|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_1/m^{1/2}.$$

Then

$$\|x_i^{\mathrm{T}}\{\widehat{\beta} - \widehat{\beta}^{(2m)}\}\|_{2,n} = \|x_i^{\mathrm{T}}\sum_{k\geq 3}\{\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\}\|_{2,n} \leq \sum_{k\geq 3}\|x_i^{\mathrm{T}}\{\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\}\|_{2,n}$$
$$\leq \{\phi^x_{\max}(m)\}^{1/2}\sum_{k\geq 3}\|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\| \leq \{\phi^x_{\max}(m)\}^{1/2}\sum_{k\geq 2}\|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_1/m^{1/2}$$
$$= \{\phi^x_{\max}(m)\}^{1/2}\|\widehat{\beta} - \widehat{\beta}^{(m)}\|_1/m^{1/2} \leq \{\phi^x_{\max}(m)\}^{1/2}\|\widehat{\beta} - \beta_0\|_1/m^{1/2},$$

where the last inequality follows from the arguments used to show the first result. □

REFERENCES

[1] Donald WK Andrews. Empirical process methods in econometrics. *Handbook of Econometrics*, 4:2247–2294, 1994.

[2] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2430, November 2012.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Ann. Statist.*, 39(1):82–130, 2011.

[4] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics: The 2010 World Congress of the Econometric Society*, 3:245–295, 2013.

[5] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.*, 81:608–650, 2014.

[6] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42:757–788, 2014.

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

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

[9] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

[10] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42:1564–1597, 2014.

[11] Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *J. Econometrics*, 142:379–398, 2008.

[12] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized Processes: Limit Theory and Statistical Applications*. Springer, New York, 2009.

[13] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *J. Multivariate Anal.*, 73(1):120–135, 2000.

[14] P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, 1:799–821, 1973.

[15] Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.*, 86(1):4–29, 2004.

[16] Roger Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

[17] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

[18] Sokbae Lee. Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory*, 19:1–31, 2003.

[19] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Econometric Theory*, 21:21–59, 2005.

[20] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

[21] Hua Liang, Suojin Wang, James M. Robins, and Raymond J. Carroll. Estimation in partially linear models with missing covariates. *J. Amer. Statist. Assoc.*, 99(466):357–367, 2004.

[22] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[23] S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.*, 12:1298–1309, 1984.
[24] S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.*, 13:1251–1638, 1985.
[25] J. L. Powell. Censored regression quantiles. *J. Econometrics*, 32:143–155, 1986.
[26] James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, 90(429):122–129, 1995.
[27] P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.
[28] Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, July 2005.
[29] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, 59:3434–3447, 2013.
[30] R. J. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.
[31] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.
[32] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, 1996.
[33] Lie Wang. L_1 penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.*, 120:135–151, 2013.
[34] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242, 2014.