Biometrika (2014), **xx**, x, *pp*. 1–17

Printed in Great Britain

Uniform Post Selection Inference for LAD Regression and Other Z-estimation Problems

BY A. BELLONI

Fuqua School of Business, Duke University,
100 Fuqua Drive, Durham, NC 27708, U.S.
abn5@duke.edu

AND V. CHERNOZHUKOV

Department of Economics, Massachusetts Institute of Technology,
52 Memorial Drive, Cambridge MA 02142, U.S.
vchern@mit.edu

AND K. KATO

Graduate School of Economics, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0013, Japan
kkato@e.u-tokyo.ac.jp

SUMMARY

We develop uniformly valid confidence regions for regression coefficients in a high-dimensional sparse least absolute deviation/median regression model. The setting is one where the number of regressors p could be large in comparison to the sample size n, but only $s \ll n$ of them are needed to accurately describe the regression function. Our new methods are based on the instrumental median regression estimator that solves the optimal estimating equation assembled from the output of the post ℓ_1-penalized median regression and post ℓ_1-penalized least squares in an auxiliary equation. The estimating equation is immunized against non-regular estimation of nuisance part of the median regression function by using Neyman's orthogonalization. We establish that in a homoscedastic regression model, the instrumental median regression estimator of a single regression coefficient is asymptotically root-n normal uniformly with respect to the underlying sparse model. The resulting confidence regions are valid uniformly with respect to the underlying model. We illustrate the value of uniformity with Monte-Carlo experiments which demonstrate that standard/naive post-selection inference breaks down over large parts of the parameter space, and the proposed method does not. We then generalize our method to the case where $p_1 \gg n$ regression coefficients are of interest in a non-smooth Huber's Z-estimation framework with approximately sparse nuisance functions, containing median regression with a single target regression coefficient as a very special case. We extend Huber's results on asymptotic normality from $p_1 \ll n$ to $p_1 \gg n$ setting, demonstrating uniform asymptotic normality over recangles, in particular, constructing simultaneous confidence bands on all p_1 coefficients and establishing their uniform validity over the underlying approximately sparse models.

Some key words: uniformly valid inference, instruments, Neymanization, optimality, sparsity, model selection

1. INTRODUCTION

We consider the following 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 of interest, whose coefficient α_0 we would like to estimate and perform (robust) inference on. The $(x_i)_{i=1}^n$ are other high-dimensional regressors or controls.

The regression error ϵ_i is independent of d_i and x_i and has median 0. The errors $(\epsilon_i)_{i=1}^n$ are independent and identically distributed with distribution function $F_\epsilon(\cdot)$ and probability density function $f_\epsilon(\cdot)$ such that $F_\epsilon(0) = 1/2$ and $f_\epsilon(0) > 0$. The assumption on the error term motivates the use of the least absolute deviation (LAD) or median regression, suitably adjusted for use in high-dimensional settings.

The dimension p of controls x_i is large, potentially much larger than n, which creates a challenge for inference on α_0. Although the unknown true 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 s can depend on n; we shall use array asymptotics). This in turn motivates the use of regularization or model selection methods.

A standard (non-robust) approach towards inference in this setting would be first to perform model selection via the ℓ_1-penalized LAD regression estimator

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} \mathbb{E}_n(|y - d\alpha - x^{\mathrm{T}}\beta|) + \frac{\lambda}{n}\|\Psi(\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}\|_1, \tag{2}$$

where λ is a penalty parameter and $\Psi^2 = \mathrm{diag}\{\mathbb{E}_n(d^2), \mathbb{E}_n(x_1^2), \ldots, \mathbb{E}_n(x_p^2)\}$ is a diagonal matrix with normalization weights, where the notation $\mathbb{E}_n(\cdot)$ denotes the average over index $1 \le i \le n$. Then, one would use the post-model selection estimator

$$(\widetilde{\alpha}, \widetilde{\beta}) \in \arg\min_{\alpha,\beta} \left\{ \mathbb{E}_n(|y - d\alpha - x^{\mathrm{T}}\beta|) : \beta_j = 0 \text{ if } \widehat{\beta}_j = 0 \right\} \tag{3}$$

to perform usual inference for α_0.

This standard approach is justified if (2) achieves perfect model selection with probability approaching 1, so that the estimator (3) has the oracle property. However conditions for perfect selection are very restrictive in this model, in particular, requiring significant 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 – which implies that usual inference breaks down and is not valid. We shall demonstrate the breakdown of such naive inference in the Monte-Carlo experiments where non-zero coefficients in β_0 are not significantly separated from zero.

The breakdown of inference does not mean that the aforementioned procedures are not suitable for prediction purposes. Indeed, the ℓ_1-LAD estimator (2) and post ℓ_1-LAD estimator (3) attain essentially optimal rates $\{(s \log p)/n\}^{1/2}$ of convergence for estimating the entire median regression function, as has been shown in Belloni & Chernozhukov (2011); Kato (2011); Wang (2013). This property means that while these procedures will not deliver perfect model recovery, they will only make moderate model selection mistakes (omitting only controls with coefficients local to zero).

To construct 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 Belloni et al. (2012), Belloni et al. (2013a), Zhang & Zhang (2014), Belloni et al. (2014a). This allows us to overcome the impact of (moderate) model selection mistakes on inference, avoiding (in part) the criticisms in Leeb & Pötscher (2005), who prove that the oracle property sometime achieved by the naive estimators necessarily implies the failure of uniform validity of inference and their semiparametric inefficiency (Leeb & Pötscher, 2008).

In order to achieve robustness with respect to moderate model selection mistakes, it will be necessary to construct orthogonal estimating equation for the target parameter. Towards that goal the following auxiliary equation plays a key role (in the homoscedastic case):

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

describing the relevant dependence of the regressor of interest d_i to the other controls x_i. We shall assume the sparsity of θ_0, namely $T_d = \mathrm{supp}(\theta_0)$ has at most $s < n$ elements, and estimate the relation (4) via Lasso or post-Lasso least squares methods described below.

Given v_i, which partials out the effect of x_i from d_i, we shall use it as an "instrument" in the following estimating equations 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 this equation to form an instrumental median regression estimator of α_0, using a plug-in estimator for $x_i^{\mathrm{T}}\beta_0$. The estimating equation above has the following orthogonality property:

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

As a result, the estimator of α_0 will be immunized against crude estimation of $x_i^{\mathrm{T}}\beta_0$, for example, via a post-selection procedure or some regularization procedure. Such orthogonalization ideas can be traced back to Neyman (1959, 1979).

Our estimation procedure has the following 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); (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, and relies on a different identification condition which in turn requires a different estimation procedure.

Step (i) constructs an estimate for the nuisance function $x_i^{\mathrm{T}}\beta_0$ and not an estimate for α_0. Here we do not need a $n^{1/2}$-rate consistency for the estimates of the nuisance function; slower rate like $o(n^{-1/4})$ will suffice. Thus, this can be based either on the ℓ_1-LAD regression estimator (2) or the associated post-model selection estimator (3).

Step (ii) estimates the residuals v_i in the decomposition (4). In order to estimate v_i we rely either on heteroscedastic Lasso (Belloni et al., 2012), a version of the Lasso estimator of Tibshirani (1996):

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n\{(d - x^{\mathrm{T}}\theta)^2\} + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1 \text{ and set } \widehat{v}_i = d_i - x_i^{\mathrm{T}}\widehat{\theta} \quad (i = 1, \ldots, n), \tag{7}$$

where λ and $\widehat{\Gamma}$ are the penalty level and data-driven penalty loadings described in Belloni et al. (2012) (restated in Appendix D), or the associated post-model selection estimator (Post-Lasso) (Belloni & Chernozhukov, 2013; Belloni et al., 2012), defined as

$$\widetilde{\theta} \in \arg\min_{\theta} \left\{\mathbb{E}_n\{(d - x^{\mathrm{T}}\theta)^2\} : \theta_j = 0 \text{ if } \widehat{\theta}_j = 0\right\} \text{ and set } \widehat{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}. \tag{8}$$

Step (iii) constructs an estimator $\check{\alpha}$ of the coefficient α_0 via an instrumental LAD regression proposed in Chernozhukov & Hansen (2008), using $(\widehat{v}_i)_{i=1}^n$ as instruments, defined formally by

$$\check{\alpha} \in \arg\min_{\alpha\in\widehat{\mathcal{A}}} L_n(\alpha),\ \text{with } L_n(\alpha) = \frac{4|\mathbb{E}_n\{\varphi(y - x^{\mathrm{T}}\widehat{\beta} - d\alpha)\widehat{v}\}|^2}{\mathbb{E}_n(\widehat{v}^2)}, \tag{9}$$

where $\varphi(t) = 1/2 - 1\{t \leq 0\}$ and $\widehat{\mathcal{A}}$ is a (possibly stochastic) parameter space for α_0. We use $\widehat{\mathcal{A}} = [\widehat{\alpha} \pm 10/(\{\mathbb{E}_n(d^2)\}^{1/2}\log n)]$, though other choices for $\widehat{\mathcal{A}}$ are possible.

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 1 and 2, the estimator $\check{\alpha}$ obeys

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

where $\sigma_n^2 = 1/\{4f_\epsilon^2 E(v^2)\}$ with $f_\epsilon = f_\epsilon(0)$ is the semi-parametric efficiency bound for regular estimators of α_0. An alternative, and more robust for practice, expression for σ_n^2 is given by Huber's sandwich:

$$\sigma_n^2 = J^{-1}\Omega J^{-1}, \text{ where } \Omega = E(v^2)/4 \text{ and } J = E(f_\epsilon dv). \tag{11}$$

We recommend to estimate Ω by the plug-in method and to estimate J by Powell's method (Powell, 1986). Furthermore, we show that the criterion function at the true value α_0 in Step 3 has the following pivotal behavior

$$nL_n(\alpha_0) \rightsquigarrow \chi^2(1). \tag{12}$$

This allows the construction of a confidence region $\widehat{A}_\xi$ with asymptotic coverage $1 - \xi$ based on the statistic L_n,

$$\text{pr}(\alpha_0 \in \widehat{A}_\xi) \to 1 - \xi \text{ where } \widehat{A}_\xi = \{\alpha \in \widehat{\mathcal{A}} : nL_n(\alpha) \leq (1-\xi)\text{-quantile of } \chi^2(1)\}. \tag{13}$$

Importantly, the robustness with respect to moderate model selection mistakes, which occurs because of (6), allows the results (10) and (12) to hold uniformly over a large range of data generating processes, similarly to the results for instrumental regression and partially linear mean regression model established in Belloni et al. (2012, 2014a). One of our proposed algorithms explicitly uses ℓ_1-regularization methods, similarly to Belloni et al. (2012) and Zhang & Zhang (2014), while the main algorithm we propose uses post-selection methods, similarly to Belloni et al. (2012, 2014a).

Throughout the paper, we use array asymptotics – asymptotics where the model changes with n – to better capture some 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 substantial regions of data-generating processes.

In Section 3 we generalize the LAD regression to a more general setting by (i) allowing p_1-dimensional target parameters defined via Huber's Z-problems are of interest, with dimension p_1 potentially much larger than the sample size n, and (ii) 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 we establish uniform asymptotic normality, using central limit theorems and bootstrap results of Chernozhukov et al. (2013), 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 prior results of Huber (1973), Portnoy (1984, 1985), He & Shao (2000) from $p_1 \ll n$ to $p_1 \gg n$.

1·1. *Notation and convention*

The notation $\mathbb{E}_n(\cdot)$ denotes the average over index $1 \leq i \leq n$, that is, it simply abbreviates $n^{-1}\sum_{i=1}^n(\cdot)$. For example, $\mathbb{E}_n(x_j^2) = n^{-1}\sum_{i=1}^n x_{ij}^2$. 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 write the support of a vector $\delta \in \mathbb{R}^p$ as $\text{supp}(\delta) = \{j \in \{1,\ldots p\} : \delta_j \neq 0\}$. We use the notation $a \vee b = \max\{a,b\}, a \wedge b = \min\{a,b\}$, and the arrow $\rightsquigarrow$ denotes convergence in distribution. Denote by $\Phi(\cdot)$ the distribution function of the standard normal distribution. We assume that the quantities such as p (the dimension of x_i), s (a bound on the numbers of non-zero elements of β_0 and θ_0), 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 $n \to \infty$. However, for the notational convenience, we shall omit the dependence of these quantities on n.

For a class of measurable functions $\mathcal{F}$, let $N(\epsilon, \mathcal{F}, \|\cdot\|_{Q,2})$ denote its ϵ-covering number with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$, where Q is finitely discrete, and let $\operatorname{ent}(\varepsilon, \mathcal{F}) = \log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2})$ 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 before uses a different identification condition. Several combinations are possible to implement each step, two of which are the following.

Algorithm 1 (Based on Post-Model Selection estimators).

1. Run Post-ℓ_1-penalized LAD (3) of y_i on d_i and x_i; keep fitted value $x_i^{\mathrm{T}}\widetilde{\beta}$.
2. Run Post-Lasso (8) of d_i on x_i; keep the residual $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}$.
3. Run Instrumental LAD regression (9) of $y_i - x_i^{\mathrm{T}}\widetilde{\beta}$ on d_i using $\widehat{v}_i$ as the instrument for d_i to compute the estimator $\check{\alpha}$. Report $\check{\alpha}$ and/or perform inference based upon (10) or (13).

Algorithm 2 (Based on Regularized Estimators).

1. Run ℓ_1-penalized LAD (2) of y_i on d_i and x_i; keep fitted value $x_i^{\mathrm{T}}\widehat{\beta}$.
2. Run Lasso of (7) d_i on x_i; keep the residual $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widehat{\theta}$.
3. Run Instrumental LAD regression (9) of $y_i - x_i^{\mathrm{T}}\widehat{\beta}$ on d_i using $\widehat{v}_i$ as the instrument for d_i to compute the estimator $\check{\alpha}$. Report $\check{\alpha}$ and/or perform inference based upon (10) or (13).

Remark 1 (*Penalty Levels*). In order to perform ℓ_1-LAD and Lasso, one has to suitably choose the penalty levels. We record our penalty choices In the Supplementary Appendix 3.

Remark 2 (*Differences*). Algorithm 1 relies on Post-ℓ_1-LAD and Post-Lasso while Algorithm 2 relies on ℓ_1-LAD and Lasso. Algorithm 1 relies on post-selection estimations that refit the non-zero coefficients without the penalty term, to reduce the bias, while Algorithm 2 relies on the penalized estimators. Step 3 of both algorithms relies on instrumental LAD regression.

Remark 3 (*Alternative Implementations*). In Step 2, Dantzig selector (Candes & Tao, 2007), square-root Lasso (Belloni et al., 2011) or the associated post-model selection could be used instead of Lasso or Post-Lasso. In step 3, we can use instead a one-step estimator from the ℓ_1-LAD estimator $\widehat{\alpha}$ of the form $\check{\alpha} = \widehat{\alpha} + [\mathbb{E}_n\{f_\epsilon(0)\widehat{v}^2\}]^{-1}\mathbb{E}_n\{\varphi(y - d\widehat{\alpha} - x^{\mathrm{T}}\widehat{\beta})\widehat{v}\}$ or a LAD regression with all the covariates selected in Steps 1 and 2.

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 $E(\widetilde{x}\widetilde{x}^{\mathrm{T}})$ with $\widetilde{x}_i = (d_i, x_i^{\mathrm{T}})^{\mathrm{T}}$ plays an important role in the analysis of ℓ_1-penalized LAD and Lasso. Define the minimal and maximal m-sparse eigenvalues of the population Gram matrix as

$$\bar{\phi}_{\min}(m) = \min_{1\le\|\delta\|_0\le m} \frac{\delta^{\mathrm{T}}E(\widetilde{x}\widetilde{x}^{\mathrm{T}})\delta}{\|\delta\|^2} \text{ and } \bar{\phi}_{\max}(m) = \max_{1\le\|\delta\|_0\le m} \frac{\delta^{\mathrm{T}}E(\widetilde{x}\widetilde{x}^{\mathrm{T}})\delta}{\|\delta\|^2}, \quad (14)$$

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

The main condition (Condition 1) contains sparsity of 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. *(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* $(\epsilon_i)_{i=1}^n$ *is a sequence of independent and identically distributed random variables with common distribution function* F_ϵ *such that* $F_\epsilon(0) = 1/2$, *independent of the random vectors* $\{(d_i, x_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$. *(ii)* $E(v^2 \mid x) \ge c_1$ *and* $E(|v|^3 \mid x) \le C_1$ *almost surely; moreover,* $E(d^4) + E(v^4) + \max_{1\le j\le p} E(x_j^2 d^2) + E(|x_j v|^3) \le C_1$. *(iii) There exists* $s = s_n \ge 1$ *such that* $\|\beta_0\|_0 \le s$ *and* $\|\theta_0\|_0 \le s$. *(iv) The error distribution* F_ϵ *is absolutely continuous with continuously differentiable density* $f_\epsilon(\cdot)$ *such that* $f_\epsilon(0) \ge c_1$ *and* $f_\epsilon(t) \vee |f'_\epsilon(t)| \le C_1$ *for all* $t \in \mathbb{R}$. *(v) There exist constants* K_n *and* M_n *such that* $K_n \ge \max_{1\le j\le p} |x_{ij}|$ *and* $M_n \ge 1 \vee |x_i^{\mathrm{T}}\theta_0|$ *almost surely, and they obey the growth condition* $\{K_n^4 + (K_n^2 \vee M_n^4)s^2 + M_n^2 s^3\} \log^3(p \vee n) \le n\delta_n$. *(vi)* $c_1 \le \bar{\phi}_{\min}(\ell_n s) \le \bar{\phi}_{\max}(\ell_n s) \le C_1$.

Remark 4. 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 the LAD literature (see Koenker, 2005) and in the instrumental quantile regression literature (Chernozhukov & Hansen, 2008). Condition 1 (v) restricts the sparsity index, so that $s^3 \log^3(p \vee n) = o(n)$ is required; this is analogous to the restriction $p^3(\log p)^2 = o(n)$ made in He & Shao (2000) in the problem without selection. The uniformly bounded regressors condition can be relaxed with minor modifications provided the bound holds with probability approaching one. Most importantly, no assumptions on the separation from zero of the non-zero coefficients of θ_0 and β_0 are made.

Remark 5. Condition 1 (vi) is quite plausible for many designs of interest. Combined with Condition 1 (v), an equivalence between the norms induced by the empirical Gram matrix and the population Gram matrix over s-sparse vectors follows. Examples of such equivalence are: Theorem 3.2 in Rudelson & Zhou (2013) for independent and identically distributed sub-Gaussian regressors and $s \log^2(n \vee p) \le \delta_n n$; Theorem 4.3 in Rudelson & Zhou (2013) for independent and identically distributed uniformly bounded regressors and $s(\log^3 n) \log(p \vee n) \le \delta_n n$.

2·3. *Results*

We begin with considering the estimators generated by Algorithms 1 and 2.

THEOREM 1 (ROBUST ESTIMATION AND INFERENCE). *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* $n \ge 1$. *Moreover, suppose that with probability at least* $1 - \Delta_n$, $\|\widehat{\beta}\|_0 \le C_1 s$. *Then, as* $n \to \infty$, *for* $\sigma_n^2 = 1/\{4 f_\epsilon^2 E(v^2)\}$,

$$\sigma_n^{-1} n^{1/2} (\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1) \ \textit{and}\ \ nL_n(\alpha_0) \rightsquigarrow \chi^2(1).$$

Theorem 1 Algorithms 1 and 2 produce estimators $\check{\alpha}$ that perform equally well, to the first order, with asymptotic variance equal to semi-parametric efficiency bound σ_n^2. Both algorithms rely on sparsity of $\widehat{\beta}$ and $\widehat{\theta}$. Sparsity of the former follows immediately under sharp penalty choices for optimal rates as shown in Supplementary Appendix 3·3. The sparsity for the latter potentially requires heavier penalty as shown in Belloni & Chernozhukov (2011); alternatively, sparsity for the estimator in Step 1 can also be achieved by truncating the smallest components of $\widehat{\beta}$. Lemma 6 in Appendix 4 shows that a suitable truncation gets 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 (**UNIFORMLY VALID CONFIDENCE INTERVALS**). *Let $\check{\alpha}$ 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),*

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

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

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 1 and 2. Condition 1 explicitly characterize regions of data-generating processes for which the uniformity result holds. Simulations results presented below also provide an 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.

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 but we need to consider the weighted version of the auxiliary equation (4) in order to achieve the semiparametric efficiency bound. The analysis of the impact of such estimation is very delicate and is developed in the companion work (Belloni et al., 2013b).

2·4. *Generalization to Many Target Coefficients with Inifinite Dimensional Nuisance Parameters*

We consider the following generalization to the previous model:

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

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

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

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

where α_j is the target coefficient,

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

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 α_j 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 (Bickel et al., 2009; Belloni et al., 2014b) with Fourier coefficients. 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 (as stated precisely in Section 3). 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 pre-determined 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 estimating equations

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

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:

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

This estimation problem is subsumed as *special case* in the next section.

3. Inference on Many Target Parameters in Z-Problems with Infinite Dimensional Nuisance Functions

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 the median regression example, its generalization discussed above, as well many other semi-parametric models.

The interest lies in $p_1 = p_{1n}$ real-valued target parameters α_j $(1 \leq j \leq p_1)$. We assume that $\alpha_j \in \mathcal{A}_j$ for every $1 \leq j \leq p_1$, 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 following moment condition:

$$E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = 0. \tag{15}$$

Here vector w is a random vector taking values in $\mathcal{W}$, a Borel subset of a Euclidean space, which contains vectors z_j $(1 \leq j \leq p_1)$ as subvectors, and each z_j takes values in $\mathcal{Z}_j$ (z_j and $z_{j'}$ with $1 \leq j \neq j' \leq p_1$ may have 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

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

where $p = p_n$ is possibly much larger than n while $s = s_n$, the sparsity level of $\widehat{h}_j$, is $\ll n$, 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 (15):

$$|\mathbb{E}_n[\psi_j\{w,\widehat{\alpha}_j,\widehat{h}_j(z_j)\}]| \le \inf_{\alpha\in\widehat{\mathcal{A}}_j} |\mathbb{E}_n[\psi\{w,\alpha,\widehat{h}_j(z_j)\}]| + \epsilon_n, \tag{16}$$

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 (immunity) of the scores with respect to small perturbations in the value of the nuisance parameters, which we can express in the following condition:

$$\partial_t E\{\psi_j(w,\alpha_j,t) \mid z_j\}|_{t=h_j(z_j)} = 0, \tag{17}$$

where we use the symbol ∂_t to abbreviate $\partial/\partial t$. It is important to construct the scores ψ_j to have property (17). Generally, we can construct the scores ψ_j that obey (17) by projecting some initial non-orthogonal scores onto the orthogonal complement of the tangent space for the nuisance parameter (see van der Vaart & Wellner, 1996; van der Vaart, 1998; Kosorok, 2008). 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 $b_n = \{\log(ep_1)\}^{1/2}$ (recall that the dependence of p_1, p on n is implicit).

Condition 2. *For every $n \ge 1$, we observe independent and identically distributed copies of $(w_i)_{i=1}^n$ of random vector w, whose law is determined by the probability measure $P \in \mathcal{P}_n$. Uniformly in $n \ge n_0, P \in \mathcal{P}_n$, and $1 \le j \le p_1$, the following conditions are satisfied. (i) The true parameter α_j obeys (15); $\widehat{\mathcal{A}}_j$ 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 $\nu \in \{\alpha, t_1,\dots,t_M\}$, $E[\sup_{\alpha_j\in\mathcal{A}_j} |\partial_\nu E[\psi_j\{w,\alpha,h_j(z_j)\} \mid z_j]|^2] \le C_1$. Moreover, there exist constants $L_{1n} \ge 1, L_{2n} \ge 1$, 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 $\nu,\nu' \in \{\alpha,t_1,\dots,t_M\}$, $\sup_{(\alpha,t)\in\mathcal{A}_j\times\mathcal{T}_j(z_j)} |\partial_\nu\partial_{\nu'}E\{\psi_j(w,\alpha,t) \mid z_j\}| \le L_{1n}$, and for every $\alpha,\alpha' \in \mathcal{A}_j, t,t' \in \mathcal{T}_j(z_j)$, $E[\{\psi_j(w,\alpha,t) - \psi_j(w,\alpha',t')\}^2 \mid z_j] \le L_{2n}(|\alpha-\alpha'|^\varsigma + \|t-t'\|^\varsigma)$. (iii) The orthogonality condition (17) holds. (iv) The following global and local identifiability conditions hold: $2|E[\psi_j\{w,\alpha,h_j(z_j)\}]| \ge |\Gamma_j(\alpha-\alpha_j)| \wedge c_1$ for all $\alpha \in \mathcal{A}_j$, where*

$$\Gamma_j = \partial_\alpha E[\psi_j\{w,\alpha_j,h_j(z_j)\}].$$

Moreover, $|\Gamma_j| \ge c_1$. (v) The second moments of scores are bounded away from zero: $E[\psi_j^2\{w,\alpha_j,h_j(z_j)\}] \ge c_1$.

The following condition uses a notion of pointwise measurable classes of functions (see van der Vaart & Wellner, 1996, p.110 for the definition).

Condition 3. *Uniformly in $n \ge n_0, P \in \mathcal{P}_n$, and $1 \le j \le p_1$, 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 $\|(\theta_{mk})_{k=1}^p\|_0 \le s$, $\widetilde{h}_{jm}(z) \in \mathcal{T}_{jm}(z)$ for all $z \in \mathcal{Z}_j$, and $E[\{\widetilde{h}_{jm}(z_j) - h_{jm}(z_j)\}^2] \le$*

$C_1 s(\log a_n)/n$, 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 $\mathrm{ent}(\varepsilon, \mathcal{F}_j) \leq C_1\{\log(e/\varepsilon) + \sum_{m=1}^{M} \mathrm{ent}(\varepsilon/C_1, \mathcal{H}_{jm})\}$. (iii) The class $\mathcal{F}_j$ has measurable envelope $F_j \geq \sup_{f\in\mathcal{F}_j} |f|$, such that $F = \max_{1\leq j\leq p_1} F_j$ obeys $E\{F^q(w)\} \leq C_1$ for some $q \geq 4$. (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 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 He & Shao (2000).

Conditions 3 (i) and (iii) require 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) median regression function and a conditional expectation function in an auxiliary equation. There are lots of conditions in the literature that imply these conditions from various primitive assumptions. For the case with $q = \infty$, condition (vi) implies the following restrictions on the sparsity indices: $(s^2 \log^3 a_n)/n \to 0$ for the case where $\varsigma = 2$ (smooth ψ_j) and $(s^3 \log^5 a_n)/n \to 0$ for the case where $\varsigma = 1$ (non-smooth ψ_j). Condition 3 (ii) is a mild condition on ψ_j – it holds for example, when ψ_j is generated by applying monotone and Lipschitz transformations to its arguments, as was the case in median regression (see van der Vaart & Wellner, 1996, for many other ways). 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).

Define

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

We are now in position to state the main theorem of this section.

THEOREM 2 (UNIFORM BAHADUR REPRESENTATION). *Under Conditions 2 and 3, uniformly in $P \in \mathcal{P}_n$, with probability $1 - o(1)$, as $n \to \infty$,*

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

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

COROLLARY 2 (UNI-DIMENSIONAL CENTRAL LIMIT THEOREM). *Under the same conditions as in Theorem 2, as $n \to \infty$,*

$$\max_{1\leq j\leq p_1} \sup_{P\in\mathcal{P}_n} \sup_{t\in\mathbb{R}} \left| \mathrm{pr}_P \left\{ n^{1/2} \sigma_j^{-1}(\widehat{\alpha}_j - \alpha_j) \leq t \right\} - \mathrm{pr}_P\{N(0,1) \leq t\} \right| = o(1).$$

This implies, in particular, that

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

provided $\max_{1\leq j\leq p_1} |\widehat{\sigma}_j - \sigma_j| = o_P(1)$ uniformly in $P \in \mathcal{P}_n$.

This result constructs pointwise confidence intervals for α_j, and shows that they are valid uniformly in $P \in \mathcal{P}_n$ and $1 \leq j \leq p_1$.

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 Chernozhukov et al. (2013) on central limit theorem for p_1-dimensional (approximate) sample means, with $p_1 \gg n$. Let

$$\mathcal{N} = (\mathcal{N}_j)_{1 \leq j \leq p_1} \sim N(0, \Omega)$$

be a random vector with normal distribution with mean zero and covariance matrix $\Omega = (E\{\phi_j(w)\phi_{j'}(w)\})_{1 \leq j,j' \leq p_1}$. 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 $A = B = \{1, \ldots, p_1\}$, $R = \{z \in \mathbb{R}^{p_1} : \max_{1 \leq j \leq p_1} |z_j| \leq t\}$.

COROLLARY 3 (HIGH-DIMENSIONAL CENTRAL LIMIT THEOREM OVER RECTANGLES). *Under the same conditions as in Theorem* 2*, provided that* $(\log p_1)^7 = o(n)$*,*

$$\sup_{P \in \mathcal{P}_n} \sup_{R \in \mathcal{R}} \left| \mathrm{pr}_P \left[n^{1/2} \{\sigma_j^{-1} (\widehat{\alpha}_j - \alpha_j)\}_{j=1}^{p_1} \in R \right] - \mathrm{pr}_P \{\mathcal{N} \in R\} \right| = o(1). \tag{18}$$

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

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

The result provides simultaneous confidence bands for $(\alpha_j)_{j=1}^{p_1}$, which are valid uniformly in $P \in \mathcal{P}_n$. Moreover, (18) is immediately useful for *multiple hypotheses testing* about $(\alpha_j)_{j=1}^{p_1}$ via the step-down methods of Romano & Wolf (2005) which control the family-wise error rates – see Chernozhukov et al. (2013) for further discussion of multiple testing with $p_1 \gg n$.

In practice the distribution of $\mathcal{N}$ is unknown due to the unknown covariance matrix, but it can be approximated by the Gaussian multiplier bootstrap, which generates a vector $\mathcal{N}^*$ as follows:

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

where $(\xi_i)_{i=1}^n$ are independent and identically distributed draws of standard normal random variables, which are independently distributed of the data $(w_i)_{i=1}^n$, and $\widehat{\phi}_j$ are any estimators of ϕ_j, such that $\max_{1 \leq j,j' \leq p_1} |\mathbb{E}_n\{\widehat{\phi}_j(w)\widehat{\phi}_{j'}(w)\} - \mathbb{E}_n\{\phi_j(w)\phi_{j'}(w)\}| = o_P(b_n^{-4})$ uniformly in $P \in \mathcal{P}_n$. Let $\widehat{\sigma}_j^2 = \mathbb{E}_n\{\widehat{\phi}_j^2(w)\}$. Theorem 3.2 in Chernozhukov et al. (2013) on multiplier bootstrap for approximate means then implies the following result.

COROLLARY 4 (VALIDITY OF MULTIPLIER BOOTSTRAP). *Under the same conditions as in 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* $\max_{1 \leq j \leq p_1} |\mathcal{N}_j^*|$*,*

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

4. MONTE-CARLO EXPERIMENTS

In this section we examine the finite sample performance of the proposed estimators. We focus on the estimator constructed by Algorithm 1, which is based on post-model selection methods.

We considered the following regression model:

$$y = d\alpha_0 + x^{\mathrm{T}}(c_y\theta_0) + \epsilon, \quad d = x^{\mathrm{T}}(c_d\theta_0) + v, \tag{20}$$

where $\alpha_0 = 1/2$, $\theta_{0j} = 1/j^2, j = 1, \ldots, 10$, and $\theta_{0j} = 0$ otherwise, $x = (1, z^{\mathrm{T}})^{\mathrm{T}}$ consists of an intercept and covariates $z \sim N(0, \Sigma)$, and the errors ϵ and v are independently and identically distributed as $N(0, 1)$. 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 $\rho = 0.5$. The coefficients c_y and c_d are used to control the R^2 of the reduce form equation. For each equation, we consider the following values for the R^2: $\{0, 0.1, 0.2, \ldots, 0.8, 0.9\}$. Therefore we have 100 different designs and results are based on 500 repetitions for each design. For each repetition we draw new vectors x_i's and errors ϵ_i's and v_i's.

The design above with $x^{\mathrm{T}}(c_y\theta_0)$ is a sparse model. However, the decay of the components of θ_0 rules out typical separation from zero assumptions of the coefficients of important covariates (since the last component is of the order of $1/n$), unless c_y is very large. Thus, we anticipate that standard post-selection inference procedures – which rely on model selection of the outcome equation only – work poorly in the simulation study. In contrast, based upon the prior theoretical arguments, we anticipate that our instrumental median estimator – which works off both equations in (20)– to work well in the simulation study.

The simulation study focuses on Algorithm 1. Standard errors are computed using the formula (11). (Algorithm 2 worked similarly, though somewhat worse due to larger biases). As the main benchmark we consider the standard post-model selection estimator $\widetilde{\alpha}$ based on the post ℓ_1-penalized LAD method, as defined in (3).

In Figure 1, we display the (empirical) rejection probability of tests of a true hypothesis $\alpha = \alpha_0$, with nominal size of tests equal to 0.05. The left-top plot shows the rejection frequency of the standard post-model selection inference procedure based upon $\widetilde{\alpha}$ (where the inference procedure assumes perfect recovery of the true model). The rejection frequency deviates very sharply from the ideal rejection frequency of 0.05. This confirms the anticipated failure (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 recovery does not happen). In sharp contrast, the right top and bottom plots show that both of our proposed procedures (based on estimator $\check{\alpha}$ and the result (10) and on the statistic L_n and the result (13)) perform well, closely tracking the ideal level of 0.05. This is achieved uniformly over all the designs considered in the study, and this confirms our theoretical results established in Corollary 1.

In Figure 2, we compare the performance of the standard post-selection estimator $\widetilde{\alpha}$ (defined in (3)) and our proposed post-selection estimator $\check{\alpha}$ (obtained via Algorithm 1). We display results in three different metrics of performance – mean bias (top row), standard deviation (middle row), and root mean square error (bottom row) of the two approaches. The significant bias for the standard post-selection procedure occurs when the indirect equation (4) is nontrivial, that is, when the main regressor is correlated to other controls. Such bias can be positive or negative depending on the particular design. The proposed post-selection estimator $\check{\alpha}$ performs well in all three metrics. The root mean square error for the proposed estimator $\check{\alpha}$ are typically much smaller than those for standard post-model selection estimators $\widetilde{\alpha}$ (as shown by bottom plots in Figure 2). This is fully consistent with our theoretical results and semiparametric efficiency of the proposed estimator.

ACKNOWLEDGMENT

We would like to thank participants of Luminy conference on Oberwolfach workshop on Frontiers in Quantile Regression (November 2012, Report No. 56/2012), Nonparametric and high-dimensional statistics (December 2012),

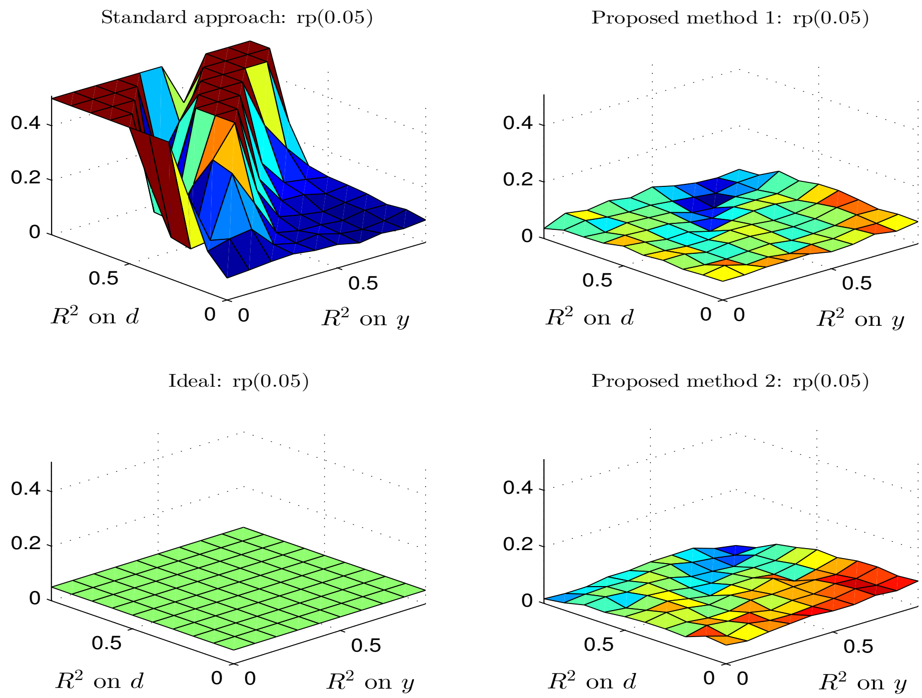

Fig. 1. The figure displays the empirical rejection probabilities of the nominal 5% level tests of a true hypothesis based on different testing procedures: the top left plot is based on the standard post-model selection procedure based on $\widetilde{\alpha}$, the top right plot is based on the proposed post-model selection procedure based on $\check{\alpha}$, and the bottom left plot is based on another proposed procedure based on the statistic L_n. Ideally we should observe the 5% rejection rate (of a true null) as in bottom left plot.

8th World Congress in Probability and Statistics (August 2012), and seminar at the University of Michigan (October 2012). We are grateful to the editor Anthony Davison, the referees, Sara van de Geer, Xuming He, Richard Nickl, Roger Koenker, Vladimir Koltchinskii, Enno Mammen, Steve Portnoy, Philippe Rigollet, and Bin Yu for useful comments and discussions.

SUPPLEMENTARY MATERIAL

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

APPENDIX 1: PROOFS FOR SECTION 3

A·1. *A Maximal Inequality*

LEMMA A1 (CHERNOZHUKOV ET AL. (2012)). *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 $\mathcal{F}$: $\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^n [f(w_i) - E\{f(w)\}], f \in \mathcal{F}$. Let $\sigma > 0$ be any positive constant such that*

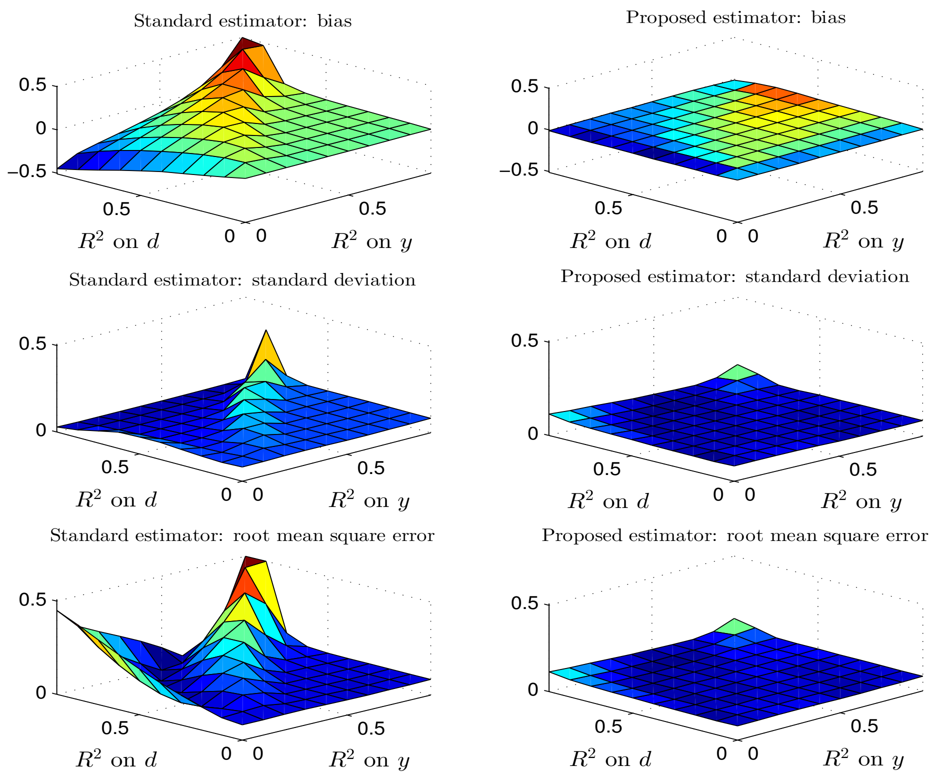

Fig. 2. The figure displays mean bias (top row), standard deviation (middle row), and root mean square error (bottom row) for the proposed post-model selection estimator $\check{\alpha}$ (right column) and the standard post-model selection estimator $\widetilde{\alpha}$ (left column).

$\sup_{f\in\mathcal{F}} E\{f^2(w)\} \leq \sigma^2 \leq E\{F^2(w)\}$. *Moreover, suppose that there exist constants* $A \geq e$ *and* $s \geq 1$ *such that* $\text{ent}(\varepsilon, \mathcal{F}) \leq s\log(A/\varepsilon)$ *for all* $0 < \varepsilon \leq 1$. *Then*

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

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

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

where K_q *is a constant that depends only on q.*

Proof. The first inequality follows from Corollary 5.1 in Chernozhukov et al. (2012). The second inequality follows from application of Theorem 5.1 in Chernozhukov et al. (2012) with $\alpha = 1$ and $[E\{\max_{1\leq i\leq n} F^2(w_i)\}]^{1/2} \leq [E\{\max_{1\leq i\leq n} F^q(w_i)\}]^{1/q} \leq n^{1/q}[E\{F^q(w)\}]^{1/q}$. □

A·2. *Proof of Theorem 2*

We begin with proving the following technical lemma. Recall $a_n = \max(p_1, p, n, e)$.

LEMMA A2. *Let* $\mathcal{F} = \{w \mapsto \psi_j\{w, \alpha, \widetilde{h}(z_j)\} : 1 \leq j \leq p_1, \alpha \in \mathcal{A}_j, \widetilde{h} \in \mathcal{H}_j \cup \{h_j\}\}$. *Then we have* $\mathrm{ent}(\varepsilon, \mathcal{F}) \leq CMs\log(a_n/\varepsilon)$ *for all* $0 < \varepsilon \leq 1$ *where* C *is a constant that depends only on* C_1.

Proof of Lemma 2. Recall the classes of functions $\mathcal{F}_j$ given in Condition 3. We first note that $\mathcal{F} = \cup_{j=1}^{p_1}\mathcal{F}_j$, so that $\mathrm{ent}(\varepsilon, \mathcal{F}) \leq \log(p_1) + \max_{1\leq j\leq p_1}\mathrm{ent}(\varepsilon, \mathcal{F}_j)$. Since each $\mathcal{H}_{jm}$ consists of p choose s VC subgraph classes with VC indices bounded by $s+2$, we have $\mathrm{ent}(\varepsilon, \mathcal{H}_{jm}) \leq Cs\log(a_n/\varepsilon)$ where C is universal, so that by Condition 3 (ii) we have $\mathrm{ent}(\varepsilon, \mathcal{F}_j) \leq C_1\{\log(e/\varepsilon) + CMs\log(C_1 a_n/\varepsilon)\}$. The desired conclusion follows from adjusting the constant C. □

Proof of Theorem 2. It suffices to prove the theorem under any sequence $P = P_n \in \mathcal{P}_n$. We shall suppress the dependency 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 $P \in \mathcal{P}_n, n$, nor $1 \leq j \leq p_1$. 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 $g : \mathcal{W} \to \mathbb{R}$, $\mathbb{G}_n(g) = n^{-1/2}\sum_{i=1}^n[g(w_i) - E\{g(w)\}]$.

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

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

uniformly in $1 \leq j \leq p_1$. Expand $\mathbb{E}_n[\psi_j\{w, \widetilde{\alpha}_j, \widehat{h}_j(z_j)\}]$ as

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

We first bound III_j. Observe that, with probability $1 - o(1)$, $\max_{1\leq j\leq p_1}|III_j| \leq n^{-1/2}\sup_{f\in\mathcal{F}'}|\mathbb{G}_n(f)|$, where $\mathcal{F}'$ is the class of functions defined by

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

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

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

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 $1 \leq m \leq M$ 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 $1 - o(1)$,

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

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

Next, we expand II_j. Pick any $\alpha \in \mathcal{A}_j$ with $|\alpha - \alpha_j| \leq C\rho_n$, $\widetilde{h} = (\widetilde{h}_m)_{m=1}^M \in \mathcal{H}_j$, and $z_j \in \mathcal{Z}_j$. Then by Taylor's theorem, there exists a pair $(\bar{\alpha}, \bar{t})$ on the line segment joining $(\alpha, \widetilde{h}(z_j))$ and $(\alpha_j, h_j(z_j))$ with

$$\begin{aligned}&E[\psi_j\{w, \alpha, \widetilde{h}(z_j)\} \mid z_j] = E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j] + \partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j](\alpha - \alpha_j) \\ &+ \textstyle\sum_{m=1}^M [\partial_{t_m}E\{\psi_j(w, \alpha_j, h_j(z_j)) \mid z_j\}]\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\} + 2^{-1}\partial_\alpha^2 E\{\psi_j(w, \bar{\alpha}, \bar{t}) \mid z_j\}(\alpha - \alpha_j)^2 \\ &\quad + 2^{-1}\textstyle\sum_{m,m'=1}^M \partial_{t_m}\partial_{t_{m'}}E\{\psi_j(w, \bar{\alpha}, \bar{t}) \mid z_j\}\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}\{\widetilde{h}_{m'}(z_j) - h_{jm'}(z_j)\} \\ &\quad + \textstyle\sum_{m=1}^M \partial_\alpha\partial_{t_m}E\{\psi_j(w, \bar{\alpha}, \bar{t}) \mid z_j\}(\alpha - \alpha_j)\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}. \qquad \text{(A1)}\end{aligned}$$

Here the third term on the right side is zero because of the orthogonality condition (17). 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, the expectation of each of the last three terms is bounded by $CL_{1n}\rho_n^2 = o(n^{-1/2}b_n^{-1})$, uniformly in $1 \le j \le p_1$. Therefore, with probability $1-o(1)$, $II_j = \Gamma_j(\widetilde{\alpha}_j - \alpha_j) + o(n^{-1/2}b_n^{-1})$, uniformly in $1 \le j \le p_1$. Combining the previous bound on III_j with this expansion leads to the desired assertion.

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

$$\max_{1\le j\le p_1} |\mathbb{E}_n[\psi_j\{w,\alpha_j,h_j(z_j)\}]| \le Cn^{-1/2}\{(\log a_n)^{1/2} + n^{-1/2+1/q}\log a_n\} \le Cn^{-1/2}\log a_n.$$

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 $1-o(1)$,

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

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

Step 3. (Preliminary rate for $\widehat{\alpha}_j$). We wish to show that with probability $1-o(1)$, $\max_{1\le j\le p_1} |\widehat{\alpha}_j - \alpha_j| \le C\rho_n$. By Step 2 and the definition of $\widehat{\alpha}_j$, with probability $1-o(1)$, we have $\max_{1\le j\le p_1} |\mathbb{E}_n[\psi_j\{w,\widehat{\alpha}_j,\widehat{h}_j(z_j)\}]| = o(n^{-1/2}b_n^{-1})$. Recall $\mathcal{F} = \{w \mapsto \psi_j\{w,\alpha,\widetilde{h}(z_j)\} : 1 \le j \le p_1, \alpha \in \mathcal{A}_j, \widetilde{h} \in \mathcal{H}_j \cup \{h_j\}\}$ given in Lemma 2. Then with probability $1-o(1)$,

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

uniformly in $1 \le j \le p_1$. Applying Lemmas 1 and 2 with $\sigma = [E\{F^2(w)\}]^{1/2} \le C$ and $t = \log n$, we have, with probability $1-o(1)$,

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

Moreover, application of the expansion (A1) 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 $1-o(1)$,

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

uniformly in $1 \le j \le p_1$, where we have used Condition 2 (ii) together with the fact that $E[\{\widetilde{h}_m(z_j) - h_{jm}(z_j)\}^2] \le C\rho_n^2$ for all $1 \le m \le M$ 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 $|\Gamma_j| \ge c_1$, implies that with probability $1-o(1)$, $|\widehat{\alpha}_j - \alpha_j| \le o(n^{-1/2}b_n^{-1}) + O(\rho_n) = O(\rho_n)$, uniformly in $1 \le j \le p_1$.

Step 4. (Uniform Bahadur representation for $\widehat{\alpha}_j$). By Steps 1 and 3, with probability $1-o(1)$,

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

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

References

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

Belloni, A. & Chernozhukov, V. (2011). ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics* **39**, 82–130.

Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**, 521–547.

Belloni, A., Chernozhukov, V. & Hansen, C. (2013a). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics: 10th World Congress of Econometric Society Vol. 3* , 245–295.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014a). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies* .

Belloni, A., Chernozhukov, V. & Kato, K. (2013b). Robust inference in high-dimensional approximately sparse quantile regression models. *arXiv:1312.7186* .

Belloni, A., Chernozhukov, V. & Wang, L. (2011). Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

Belloni, A., Chernozhukov, V. & Wang, L. (2014b). Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics* .

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Candes, E. & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2313–2351.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2012). Gaussian approximation of suprema of empirical processes. *arXiv:1212.6885* .

Chernozhukov, V., Chetverikov, D. & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics* **41**, 2786–2819.

Chernozhukov, V. & Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics* **142**, 379–398.

He, X. & Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis* **73**, 120–135.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics* **1**, 799–821.

Kato, K. (2011). Group Lasso for high dimensional sparse quantile regression models. *arXiv:1103.1458* .

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory* **21**, 21–59.

Leeb, H. & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142**, 201–211.

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

Neyman, J. (1979). $C(\alpha)$ tests and their use. *Sankhya* **41**, 1–21.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics* **12**, 1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Annals of Statistics* **13**, 1251–1638.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics* **32**, 143–155.

Romano, J. P. & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica* **73**, 1237–1282.

Rudelson, M. & Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59**, 3434–3447.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society Series B* **58**, 267–288.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Wang, L. (2013). L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120**, 135–151.

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of Royal Statistical Society Series B* **76**, 217–242.

Biometrika (2014), **xx**, x, *pp*. 1–15

Printed in Great Britain

Suplementary Material for Uniform Post Selection Inference for LAD Regression and Other Z-estimation Problems

BY A. BELLONI

Fuqua School of Business, Duke University,
100 Fuqua Drive, Durham, NC 27708, U.S.
abn5@duke.edu

AND V. CHERNOZHUKOV

Department of Economics, Massachusetts Institute of Technology,
52 Memorial Drive, Cambridge MA 02142, U.S.
vchern@mit.edu

AND K. KATO

Graduate School of Economics, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0013, Japan
kkato@e.u-tokyo.ac.jp

SUMMARY

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

Some key words: uniformly valid inference, instruments, Neymanization, optimality, sparsity, model selection

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 $\|z\|_{2,n} = \{\mathbb{E}_n(z^2)\}^{1/2} = (n^{-1}\sum_{i=1}^n z_i^2)^{1/2}$. For example, for a vector $\delta \in \mathbb{R}^p$ and p-dimensional regressors $(x_i)_{i=1}^n$, $\|x^T\delta\|_{2,n} = [\mathbb{E}_n\{(x^{\mathrm{T}}\delta)^2\}]^{1/2}$ denotes the empirical prediction norm of δ. 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 $a = O_P(b)$.

1. GENERALIZATION AND ADDITIONAL RESULTS FOR THE LAD MODEL

1·1. *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. In order to achieve the semiparametric efficiency bound we need to consider the weighted version of the auxiliary equation (4). Specifically, we can rely on the following of weighted decomposition:

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

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

$$f_i = f_{\epsilon_i}(0 \mid d_i, x_i) \quad (i = 1, \ldots, n), \tag{2}$$

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′ (Based on Post-Model Selection estimators).

1. Run Post-ℓ_1-penalized LAD of y_i on d_i and x_i; keep fitted value $x_i^{\mathrm{T}}\widetilde{\beta}$.
2. Run Post-Lasso of $f_i d_i$ on $f_i x_i$; keep the residual $\widehat{v}_i^* = f_i(d_i - x_i^{\mathrm{T}}\widetilde{\theta})$.
3. Run Instrumental LAD regression of $y_i - x_i^{\mathrm{T}}\widetilde{\beta}$ on d_i using $\widehat{v}_i^*$ as the instrument for d_i to compute the estimator $\check{\alpha}$. Report $\check{\alpha}$ and/or perform inference.

An analogous generalization of Algorithm 2 based on regularized estimator results from removing the word Post in the algorithm above.

Under similar regularity conditions, uniformly over a large collection $\mathcal{P}_n^*$ of distributions of $\{(y_i, d_i, x_i^{\mathrm{T}})'\}_{i=1}^n$, the estimator $\check{\alpha}$ above obeys

$$\{4E(v^{*2})\}^{1/2}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0, 1). \tag{3}$$

Moreover, the criterion function at the true value α_0 in Step 3 also has a pivotal behavior, namely

$$nL_n(\alpha_0) \rightsquigarrow \chi^2(1), \tag{4}$$

which can also be used to construct a confidence region $\widehat{A}_{n,\xi}$ based on the L_n-statistic as in (12) 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 Belloni et al. (2013), which considers the more general problem of uniformly valid inference for quantile regression models in approximately sparse models.

1·2. *Connection to Neymanization*

In this section we make some connections to Neyman's $C(\alpha)$ test (Neyman, 1959, 1979). For the sake of exposition we assume that $(y_i, d_i, x_i)_{i=1}^n$ are independent and identically distributed but we shall use the heteroscedastic setup introduced in the previous section. We consider the estimating equation for α_0:

$$E\{\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta_0)v_i\} = 0.$$

Our problem is to find useful instruments v_i such that

$$\frac{\partial}{\partial\beta} E\{\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta)v_i\}|_{\beta=\beta_0} = 0.$$

If this property holds, the estimator of α_0 will be immunized against crude or nonregular estimation of β_0, for example, via a post-selection procedure or some regularization procedure. Such immunization ideas are in fact behind Neyman's classical construction of his $C(\alpha)$ test, so we shall use the term Neymanization to describe such procedure. There will be many instruments v_i that can achieve the property stated above, and there will be one that is optimal.

The instruments can be constructed by taking $v_i = z_i/f_i$, where z_i is the residual in the regression equation:

$$w_i d_i = w_i m_0(x_i) + z_i, \; E(w_i z_i \mid x_i) = 0, \tag{5}$$

where w_i is a nonnegative weight, a function of (d_i, z_i) only, for example $w_i = 1$ or $w_i = f_i$ (the latter choice will in fact be optimal). The function $m_0(x_i)$ solves the least squares problem

$$\min_{h \in \mathcal{H}} E[\{wd - wh(x)\}^2],$$

where $\mathcal{H}$ is the class of measurable functions $h(x)$ such that $E[w^2 h^2(x)] < \infty$. Our assumption is that the $m_0(x)$ is a sparse function $x^{\mathrm{T}}\theta_0$ with $\|\theta_0\|_0 \leq s$, so that

$$w_i d_i = w_i x_i^{\mathrm{T}} \theta_0 + z_i, \;\; E(w_i z_i \mid x_i) = 0.$$

In finite samples, the sparsity assumption allows to employ post-Lasso and Lasso to solve the least squares problem above approximately, and estimate z_i. Of course, the use of other structured assumptions may motivate the use of other regularization methods.

Arguments similar to those in the proofs show that, for $\sqrt{n}(\alpha - \alpha_0) = O(1)$,

$$\sqrt{n}\left[\mathbb{E}_n\{\varphi(y - d\alpha - x^{\mathrm{T}}\widehat{\beta})v\} - \mathbb{E}_n\{\varphi(y - d\alpha - x^{\mathrm{T}}\beta_0)v\}\right] = o_P(1),$$

for $\widehat{\beta}$ based on a sparse estimation procedure, despite the fact that $\widehat{\beta}$ converges to β_0 at a slower rate than $1/\sqrt{n}$. That is, the empirical estimating equations behave as if β_0 is known. Hence for estimation we can use $\widehat{\alpha}$ as a minimizer of the statistic:

$$L_n(\alpha) = c_n^{-1}|\sqrt{n}\mathbb{E}_n\{\varphi(y - d\alpha - x^{\mathrm{T}}\widehat{\beta})v\}|^2,$$

where $c_n = \mathbb{E}_n(v^2)/4$. Since $L_n(\alpha_0) \rightsquigarrow \chi^2(1)$, we can also use the statistic directly for testing hypotheses and for construction of confidence intervals.

This is in fact a version of Neyman's $C(\alpha)$ test statistic, adapted to the present non-smooth setting. The usual expression of $C(\alpha)$ statistic is different. To see a more familiar form, let $\theta_0 = \{E(w^2 x x^{\mathrm{T}})\}^{-} E(w^2 d x^{\mathrm{T}})$, where A^- denotes a generalized inverse of A, and write

$$v_i = (w_i/f_i)d_i - (w_i/f_i)x_i^{\mathrm{T}}\{E(w^2 x x^{\mathrm{T}})\}^{-}E(w^2 d x'), \;\text{ and }\; \widehat{\varphi}_i = \varphi(y_i - d_i\alpha - x_i'\widehat{\beta}),$$

so that,

$$L_n(\alpha) = c_n^{-1}\left|\sqrt{n}\left[\mathbb{E}_n\{\widehat{\varphi}(w/f)d\} - \mathbb{E}_n\{\widehat{\varphi}(w/f)x^{\mathrm{T}}\}\{E(w^2 x x^{\mathrm{T}})\}^{-}E(w^2 d x^{\mathrm{T}})\right]\right|^2.$$

This is indeed a familiar form of a $C(\alpha)$ statistic.

The estimator $\widehat{\alpha}$ that minimizes $L_n(\alpha)$ up to $o_P(1)$, under suitable regularity conditions, obeys

$$\sigma_n^{-1}\sqrt{n}(\widehat{\alpha} - \alpha_0) \rightsquigarrow N(0,1), \text{ where } \sigma_n^2 = \frac{1}{4}\{E(fdv)\}^{-2}E(v^2).$$

It is easy to show that the smallest value of σ_n^2 is achieved by using $v_i = v_i^*$ induced by setting $w_i = f_i$:

$$\sigma_n^{*2} = \frac{1}{4}\{E(v^{*2})\}^{-1}. \tag{6}$$

Thus, setting $w_i = f_i$ gives an optimal instrument amongst all immunizing instruments generated by the process described above. Obviously, this improvement translates into shorter confidence intervals and better testing based on either $\widehat{\alpha}$ or $L_n(\alpha)$. While $w_i = f_i$ is optimal, f_i will have to be estimated in practice, resulting actually in more stringent condition than when using non-optimal, known weights, for example, $w_i = 1$. The use of known weights may also give better behavior under misspecification of the model. Under homoscedasticity, $w_i = 1$ is an optimal weight.

1·3. *Minimax Efficiency*

There is also a clean connection to the (local) minimax efficiency analysis from the semiparametric efficiency literature. Lee (2003) derives an efficient score function for the partially linear median regression model:

$$S_i = 2\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta_0) f_i\{d_i - m_0^*(x)\},$$

where $m_0^*(x_i)$ is $m_0(x_i)$ in (5) induced by the weight $w_i = f_i$:

$$m_0^*(x_i) = \frac{E(f_i^2 d_i \mid x_i)}{E(f_i^2 \mid x_i)}.$$

Using the assumption $m_0^*(x_i) = x_i^{\mathrm{T}}\theta_0^*$, where $\|\theta_0^*\|_0 \leq s \ll n$ is sparse, we have that

$$S_i = 2\varphi(y_i - d_i\alpha_0 - x_i^{\mathrm{T}}\beta_0) v_i^*,$$

which is the score that was constructed using Neymanization. It follows that the estimator based on the instrument v_i^* is actually efficient in the minimax sense (see Theorem 18.4 in Kosorok, 2008), and inference about α_0 based on this estimator provides best minimax power against local alternatives (see Theorem 18.12 in Kosorok, 2008).

The claim above is formal as long as, given a law P_n, the least favorable submodels are permitted as deviations that lie within the overall model. Specifically, given a law P_n, we shall need to allow for a certain neighborhood $\mathcal{P}_n^\delta$ of P_n such that $P_n \in \mathcal{P}_n^\delta \subset \mathcal{P}_n$, where the overall model $\mathcal{P}_n$ is defined similarly as before, except now permitting heteroscedasticity (or we can keep homoscedasticity $f_i = f_\epsilon$ to maintain formality). To allow for this we consider a collection of models indexed by a parameter $t = (t_1, t_2)$:

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

where $\|\beta_0\|_0 \vee \|\theta_0^*\|_0 \leq s/2$ and conditions as in Section 2 hold. The case with $t = 0$ generates the model P_n; by varying t within δ-ball, we generate models $\mathcal{P}_n^\delta$, containing the least favorable deviations. By Lee (2003), the efficient score for the model given above is S_i, so we cannot have a better regular estimator than the estimator whose influence function is $J^{-1}S_i$, where $J = E(S_i^2)$. Since our model $\mathcal{P}_n$ contains $\mathcal{P}_n^\delta$, all the formal conclusions about (local minimax) optimality of our estimators hold from theorems cited above (using subsequence arguments to handle models changing with n). Our estimators are regular, since under models with $t = (O(1/\sqrt{n}), o(1))$, their first order asymptotics do not change, as a consequence of Theorem 1 in Section 2, though our theorems actually prove more than this.

1·4. *Alternative Implementation via Double Selection*

An alternative proposal for the method is reminiscent of the double selection method proposed in Belloni et al. (2014) for partial linear models. This version replaces Step 3 with a LAD regression of y on d and all covariates selected in Steps 1 and 2 (that is, the union of the selected sets). The method is described as follows:

Algoritm 3. (A Double Selection Method)

Step 1: Run Post-ℓ_1-LAD of y_i on d_i and x_i:

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} \mathbb{E}_n(|y - d\alpha - x^{\mathrm{T}}\beta|) + \frac{\lambda_1}{n}\|\Psi(\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}\|_1.$$

Step 2: Run Heteroscedastic Lasso of d_i on x_i:

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n\{(d - x^{\mathrm{T}}\theta)^2\} + \frac{\lambda_2}{n}\|\widehat{\Gamma}\theta\|_1.$$

Step 3: Run LAD regression of y_i on d_i and the covariates selected in Step 1 and 2:

$$(\check{\alpha}, \check{\beta}) \in \arg\min_{\alpha,\beta} \left\{\mathbb{E}_n(|y - d\alpha - x^{\mathrm{T}}\beta|) : \ \mathrm{supp}(\beta) \subseteq \mathrm{supp}(\widehat{\beta}) \cup \mathrm{supp}(\widehat{\theta})\right\}.$$

The double selection algorithm has three steps: (1) select covariates based on the standard ℓ_1-LAD regression, (2) select covariates based on heteroscedastic Lasso of the treatment equation, and (3) run a LAD 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 Step 1 and 2: $\widehat{T}^* = \mathrm{supp}(\widehat{\beta}) \cup \mathrm{supp}(\widehat{\theta})$. By the first order conditions for $(\check{\alpha}, \check{\beta})$ we have

$$\left\|\mathbb{E}_n\left\{\varphi(y - d\check{\alpha} - x^{\mathrm{T}}\check{\beta})(d, x^{\mathrm{T}}_{\widehat{T}^*})^{\mathrm{T}}\right\}\right\| = O\{(\max_{1\le i\le n}|d_i| + K_n|\widehat{T}^*|^{1/2})(1 + |\widehat{T}^*|)/n\},$$

which creates an orthogonal relation to any linear combination of $(d_i, x^{\mathrm{T}}_{i\widehat{T}^*})^{\mathrm{T}}$. In particular, by taking the linear combination $(d_i, x^{\mathrm{T}}_{i\widehat{T}^*})(1, -\widetilde{\theta}^{\mathrm{T}}_{\widehat{T}^*})^{\mathrm{T}} = d_i - x^{\mathrm{T}}_{i\widehat{T}^*}\widetilde{\theta}_{\widehat{T}^*} = d_i - x_i^{\mathrm{T}}\widetilde{\theta} = \widehat{v}_i$, which is the instrument in Step 2 of Algorithm 1, we have

$$\mathbb{E}_n\{\varphi(y - d\check{\alpha} - x^{\mathrm{T}}\check{\beta})\widehat{z}\} = O\{\|(1, -\widetilde{\theta}^{\mathrm{T}})^{\mathrm{T}}\|(\max_{1\le i\le n}|d_i| + K_n|\widehat{T}^*|^{1/2})(1 + |\widehat{T}^*|)/n\}.$$

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

$$\widetilde{L}_n(\alpha) = \frac{|\mathbb{E}_n\{\varphi(y - d\alpha - x^{\mathrm{T}}\check{\beta})\widehat{v}\}|^2}{\mathbb{E}_n[\{\varphi(y - d\check{\alpha} - x^{\mathrm{T}}\check{\beta})\}^2\widehat{v}^2]},$$

where $\widehat{v}_i$ is the instrument created by Step 2 of Algorithm 1. Thus the double selection estimator can be seen as an iterated version of the method based on instruments where the Step 1 estimate $\widetilde{\beta}$ is updated with $\check{\beta}$.

2. Proofs for Section 2

2·1. *Proof of Theorem* 1

The proof of Theorem 1 verifies Conditions 2 and 3 and applies Theorem 2. We will collect the properties of Post-ℓ_1-LAD and Post-Lasso together with required regularity conditions in Appendix 3. Moreover, we will use some auxiliary technical lemmas stated in Appendix 4. The proof focuses on Algorithm 1. We provide the minor adjustments for the proof for Algorithm 2 later since it is basically the same proof.

In Theorem 2, take $p_1 = 1, z = x, w = (y, d, x^{\mathrm{T}})^{\mathrm{T}}, M = 2, \psi(w, \alpha, t) = \{1/2 - 1(y \le \alpha d + t_1)\}(d - t_2), h(z) = (x^{\mathrm{T}}\beta_0, x^{\mathrm{T}}\theta_0)^{\mathrm{T}} = (g(x), m(x))^{\mathrm{T}} = h(x)$ (say), $\mathcal{A} = [\alpha_0 - c_2, \alpha_0 + c_2]$ 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: (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 $1 - o(1)$.

(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 $\alpha \in \mathcal{A}$,

$$|\partial_\nu E[\psi\{w, \alpha, h(x)\} \mid x]| \leq C_1 E(|dv| \mid x) \vee C_1 E(|v| \mid x) \vee 1.$$

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

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

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

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

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

$$\begin{aligned}
&E[\{\psi(w, \alpha, t) - \psi(w, \alpha', t')\}^2 \mid x] \leq C_1 E\{|d(d - t_2)| \mid c\}|\alpha - \alpha'| \\
&\quad + C_1 E\{|(d - t_2)| \mid x\}|t_1 - t'_1| + (t_2 - t'_2)^2 \leq C' M_n(|\alpha - \alpha'| + |t_1 - t'_1|) + (t_2 - t'_2)^2,
\end{aligned}$$

where C' is a constant depending only on c_3, C_1. Here as $|t_2 - t'_2| \leq |t_2 - m(x)| + |m(x) - t_2| \leq 2c_3$, the right side is bounded by $\sqrt{2}(C'M_n + 2c_3)(|\alpha - \alpha'| + \|t - t'\|)$. Hence we can take $L_{2n} = \sqrt{2}(C'M_n + 2c_3)$ and $\varsigma = 1$.

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

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

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

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

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

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

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

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

Verification of Condition 3: Note here that $a_n = p \vee n$ and $b_n = 1$. Next we 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 property.

The estimator $\widetilde{\beta}$ is based on Post-ℓ_1-penalized LAD with penalty parameters as suggested in Section 3·2. By assumption in Theorem 1, with probability $1 - \Delta_n$ we have $\widehat{s} = \|\widetilde{\beta}\|_0 \leq C_1 s$. Next we verify that Condition PLAD in Appendix 3 is implied by Condition 1 and invoke Lemmas 1 and 2. The assumptions on the error density $f_\epsilon(\cdot)$ in Condition PLAD (i) are assumed in Condition 1 (iv). Because of Condition 1 (v) and (vi), $\bar{\kappa}_{c_0}$ is bounded away from zero for n sufficiently large (see Bickel et al., 2009, Lemma 4.1) and $c_1 \leq \bar{\phi}_{\min}(1) \leq E(\widetilde{x}_j^2) \leq \bar{\phi}_{\max}(1) \leq C_1$ for every $1 \leq j \leq p$. Moreover, under Condition 1, by Lemma 7 we have $\max_{1\leq j\leq p+1} |\mathbb{E}_n(\widetilde{x}_j^2)/E(\widetilde{x}_j^2) - 1| \leq 1/2$ and $\phi_{\max}(\ell'_n s) \leq 2\mathbb{E}_n(d^2) + 2\phi^x_{\max}(\ell'_n s) \leq 5C_1$ with probability $1 - o(1)$ for some $\ell'_n \to \infty$. The required side condition of Lemma 1 is satisfied by relations (7) and (8) ahead. By Lemma 2 in Appendix 3 we have $\|x^{\mathrm{T}}(\widetilde{\beta} - \beta_0)\|_{P,2} \lesssim_P \sqrt{s\log(n \vee p)/n}$ 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 $E(|d|^3) = O(1)$, we have

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

Therefore, since $K_n^2 s^2 \log^2(p \vee n) \leq \delta_n n$ and $\lambda \lesssim \sqrt{n\log(p \vee n)}$ we have

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

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 1 and Remark 2 we have $|\widehat{\alpha} - \alpha_0| \lesssim \sqrt{s\log(p \vee n)/n} = o(1/\log n)$ with probability $1 - o(1)$ under $s^3 \log^3(p \vee n) \leq \delta_n n$.

The estimator $\widetilde{\theta}$ is based on Post-Lasso with penalty parameters as suggested in Section 3·3. We verify that Condition HL in Appendix 3 is implied by Condition 1 and invoke Lemma 4. Indeed, Condition HL (ii) is implied by Conditions 1 (ii) and (iv) (condition (iv) is used to ensure $\min_{1\leq j\leq p} E(x_j^2) \geq c_1$). Next since $\max_{1\leq j\leq p} E(|x_j v|^3) \leq C_1$, Condition HL (iii) is satisfied if $\sqrt{\log(p \vee n)} = o(n^{1/6})$, which is implied by Condition 1 (v). Condition HL (iv) follows from Lemma 5 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ under the condition that $K_n^4 \log p \leq \delta_n n$ and $K_n^2 s \log(p \vee n) \leq \delta_n n$. Condition HL (v) follows from Lemma 7. By Lemma 4 in Appendix 3 we have $\|x^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n}$ and $\|\widetilde{\theta}\|_0 \lesssim s$ with probability $1 - o(1)$. Thus, by Lemma 7, we have $\|x^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{P,2} \lesssim_P \sqrt{s\log(n \vee p)/n}$.

Combining the results above, we have that $\widehat{h} \in \mathcal{H} = \times_{m=1}^2 \mathcal{H}_m$ with probability $1 - o(1)$ where $\mathcal{H}_m = \{\widetilde{h}_m : \mathbb{R}^p \to \mathbb{R} : \widetilde{h}_m(x) = x^{\mathrm{T}}\theta_m, \|\theta_m\|_0 \leq C_3 s, E[\{\widetilde{h}_m(x) - h_m(x)\}^2] \leq C'_3 \ell'_n s(\log a_n)/n\}$ and $\ell'_n \uparrow \infty$ sufficiently slowly.

To verify Condition 3 (iii) note that $\mathcal{F} = \varphi(\mathcal{G}) \cdot \mathcal{H}_2$, where $\varphi(u) = 1/2 - 1(u \leq 0)$ and $\mathcal{G} = \{(y, d, x^{\mathrm{T}})^{\mathrm{T}} \mapsto y - \alpha d - h(x) : \alpha \in \mathcal{A}, h \in \mathcal{H}_1\}$. $\mathcal{H}_1$ and $\mathcal{H}_2$ are the union of $\binom{p}{C_3 s}$ VC-subgraph classes. Since φ is monotone, by Lemma 2.6.18 in van der Vaart & Wellner (1996), $\varphi(\mathcal{G})$ is also a VC-subgraph class with the same VC index. Finally, the entropy of $\mathcal{F}$ associated with the product between $\varphi(\mathcal{G})$ and $\mathcal{H}_2$ satisfies the stated entropy condition; see the proof of Theorem 3 in Andrews (1994), relation (A.7).

To verify Condition 3 (v), take $s_n = \ell_n' s$ and $\rho_n = n^{-1/2}(\sqrt{s_n \log a_n} + n^{-1/2} s_n n^{1/q} \log a_n) \lesssim n^{-1/2}\sqrt{s_n \log a_n}$ under $q \geq 4$ and $s_n^2/n = o(1)$. For $\varsigma = 1$, $L_{1n} \lesssim M_n^2$ and $L_{2n} \lesssim M_n$, the condition (18) holds provided $n^{-1}M_n^2 s_n^3 \log^3 a_n = o(1)$ and $n^{-1}M_n^4 s_n^2 \log^2 a_n = o(1)$ which are implied by Condition 1 (with ℓ_n' diverging slow enough).

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

Next we prove the second result regarding $nL_n(\alpha_0)$. First consider the denominator of $L_n(\alpha_0)$. We have that with probability $1 - o(1)$

$$\begin{aligned} |\mathbb{E}_n(\widehat{v}^2) - \mathbb{E}_n(v^2)| &= |\mathbb{E}_n\{(\widehat{v} - v)(\widehat{v} + v)\}| \leq \|\widehat{v} - v\|_{2,n}\|\widehat{v} + v\|_{2,n} \\ &\leq \|x^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n}\{2\|v\|_{2,n} + \|x^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n}\} \lesssim \delta_n, \end{aligned}$$

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

Second consider the numerator of $L_n(\alpha_0)$. Since $E[\psi\{w, \alpha_0, h(x)\}] = 0$ we have with probability $1 - o(1)$

$$\mathbb{E}_n[\psi\{w, \alpha_0, \widehat{h}(x)\}] = \mathbb{E}_n[\psi\{w, \alpha_0, h(x)\}] + o(\delta_n n^{-1/2}),$$

using representation in the displayed equation of Step 4 in the proof of Theorem 2 evaluated at α_0 instead of $\widehat{\alpha}_j$. Therefore, using the identity that $nA_n^2 = nB_n^2 + n(A_n - B_n)^2 + 2nB_n(A_n - B_n)$ with

$$A_n = \mathbb{E}_n[\psi\{w, \alpha_0, \widehat{h}(x)\}] \text{ and } B_n = \mathbb{E}_n[\psi\{w, \alpha_0, h(x)\}] \lesssim_P \{E(v^2)\}^{1/2} n^{-1/2},$$

we have

$$nL_n(\alpha_0) = \frac{4n|\mathbb{E}_n[\psi\{w, \alpha_0, \widehat{h}(x)\}]|^2}{\mathbb{E}_n(\widehat{v}^2)} = \frac{4n|\mathbb{E}_n[\psi\{w, \alpha_0, h(x)\}]|^2}{\mathbb{E}_n[\psi^2\{w, \alpha_0, h(x)\}]} + O_P(\delta_n)$$

since $E(v^2)$ is bounded away from zero. By Theorem 7.1 in de la Peña et al. (2009), 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

$$I = \frac{2\sqrt{n}\mathbb{E}_n[\psi\{w, \alpha_0, h(x)\}]}{(\mathbb{E}_n[\psi^2\{w, \alpha_0, h(x)\}])^{1/2}} \rightsquigarrow N(0,1),$$

and the desired result follows since $nL_n(\alpha_0) = I^2 + O_P(\delta_n)$.

Remark 1 (*On one-step procedure*). An inspection of the proof leads to the following stochastic expansion:

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

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

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

has the following stochastic expansion:

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

so that $\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0, 1)$.

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

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

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

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

because $s^3\log^3(n \vee p) \leq \delta_n n$ 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 $E(|d|^3) = O(1)$, we have

$$\begin{aligned}\inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}^{\mathrm{T}}\delta|^3)} &\geq \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^2\|\delta_T\|\bar{\kappa}_{c_0}}{4E(|x'\delta_x|^3)+4E(|d\delta_d|^3)} \\ &\geq \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^2\|\delta_T\|\bar{\kappa}_{c_0}}{4K_n\|\delta_x\|_1E(|x^{\mathrm{T}}\delta_x|^2)+4|\delta_d|^3E(|d|^3)} \\ &\geq \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^2\|\delta_T\|\bar{\kappa}_{c_0}}{\{4K_n\|\delta_x\|_1+4|\delta_d|E(|d|^3)/E(|d|^2)\}\{E(|x^{\mathrm{T}}\delta_x|^2)+E(|\delta_d d|^2)\}} \\ &\geq \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^2\|\delta_T\|\bar{\kappa}_{c_0}}{8(1+3c_0')\|\delta_T\|_1\{K_n+O(1)\}\{2E(|\widetilde{x}^{\mathrm{T}}\delta_x|^2)+3E(|\delta_d d|^2)\}} \\ &\geq \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^2\|\delta_T\|\bar{\kappa}_{c_0}}{8(1+3c_0')\|\delta_T\|_1\{K_n+O(1)\}E(|\widetilde{x}^{\mathrm{T}}\delta_x|^2)(2+3/\bar{\kappa}_{c_0}^2)} \\ &\geq \frac{\bar{\kappa}_{c_0}/\sqrt{s}}{8\{K_n+O(1)\}(1+3c_0')\{2+3E(d^2)/\bar{\kappa}_{c_0}^2\}} \gtrsim \frac{1}{\sqrt{s}K_n}.\end{aligned} \tag{7}$$

Therefore, since $\lambda \lesssim \sqrt{n\log(p \vee n)}$ we have

$$\frac{\sqrt{n}\bar{\kappa}_{c_0}}{\sqrt{s\log(p \vee n)}} \inf_{\delta\in\Delta_{c_0}} \frac{\|\widetilde{x}^{\mathrm{T}}\delta\|_{P,2}^3}{E(|\widetilde{x}^{\mathrm{T}}\delta|^3)} \gtrsim \frac{\sqrt{n}}{K_n s\sqrt{\log(p \vee n)}} \to \infty \tag{8}$$

under $K_n^2 s^2\log^2(p \vee n) \leq \delta_n n$.

The estimator $\widehat{\theta}$ is based on Lasso. Condition HL is implied by Condition 1 and Lemma 5 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ under the condition that $K_n^4\log p \leq \delta_n n$. By Lemma 3 we have $\|x^{\mathrm{T}}(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n}$. Moreover, by Lemma 4 we have $\|\widehat{\theta}\|_0 \lesssim s$ with probability $1 - o(1)$. The required rate in the $\|\cdot\|_{P,2}$ norm follows from Lemma 7.

3. Auxiliary Results for ℓ_1-LAD and Heteroscedastic Lasso

3·1. *Notation*

In this section we state relevant theoretical results on the performance of the estimators: ℓ_1-LAD, Post-ℓ_1-LAD, heteroscedastic Lasso, and heteroscedastic Post-Lasso. There results were developed in Belloni & Chernozhukov (2011) and Belloni et al. (2012). 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 (slack) constant chosen by users (we suggest to take $c_0 = 1.1$ but the analysis is not restricted to this choice). Moreover, let $c_0' = (c_0+1)/(c_0-1)$. Also recall the definition of the minimal and maximal m-sparse eigenvalues of a matrix A as

$$\phi_{\min}(m, A) = \min_{1\le\|\delta\|_0\le m} \frac{\delta^{\mathrm{T}} A\delta}{\|\delta\|^2} \text{ and } \phi_{\max}(m, A) = \max_{1\le\|\delta\|_0\le m} \frac{\delta^{\mathrm{T}} A\delta}{\|\delta\|^2}, \tag{9}$$

where $1 \le m \le p$. Finally, define $\phi_{\min}(m) = \phi_{\min}\{m, \mathbb{E}_n(\widetilde{x}\widetilde{x}^{\mathrm{T}})\}$, $\bar{\phi}_{\min}(m) = \phi_{\min}\{m, E(\widetilde{x}\widetilde{x}^{\mathrm{T}})\}$, $\bar{\phi}_{\max}(m) = \phi_{\max}\{m, E(\widetilde{x}\widetilde{x}^{\mathrm{T}})\}$, $\phi^x_{\min}(m) = \phi_{\min}\{m, \mathbb{E}_n(xx^{\mathrm{T}})\}$, and $\phi^x_{\max}(m) = \phi_{\max}\{m, \mathbb{E}_n(xx^{\mathrm{T}})\}$. Observe that $\phi_{\max}(m) \le 2\mathbb{E}_n(d^2) + 2\phi^x_{\max}(m)$.

3·2. ℓ_1-*Penalized LAD*

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

$$\mathrm{pr}(y_i \le \widetilde{x}_i^{\mathrm{T}}\eta_0 \mid \widetilde{x}_i) = 1/2.$$

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

$$\widehat{\eta} \in \arg\min_{\eta} \mathbb{E}_n(|y - \widetilde{x}^{\mathrm{T}}\eta|) + \frac{\lambda}{n}\|\Psi\eta\|_1,$$

where $\Psi^2 = \mathrm{diag}\{\mathbb{E}_n(\widetilde{x}_1^2), \ldots, \mathbb{E}_n(\widetilde{x}_p^2)\}$ is a diagonal matrix of penalty loadings. As established in Belloni & Chernozhukov (2011) and Wang (2013), under the event that

$$\frac{\lambda}{n} \ge 2c_0\|\Psi^{-1}\mathbb{E}_n[\{1/2 - 1(y \le \widetilde{x}^{\mathrm{T}}\eta_0)\}\widetilde{x}]\|_\infty, \tag{10}$$

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 (10) holds with high probability. In particular, the pivotal rule discussed in Belloni & Chernozhukov (2011) proposes to set $\lambda = c_0 n\Lambda(1-\gamma \mid \widetilde{x})$ with $\gamma \to 0$ where

$$\Lambda(1-\gamma \mid \widetilde{x}) = (1-\gamma)\text{-quantile of } 2\|\Psi^{-1}\mathbb{E}_n[\{1/2 - 1(U \le 1/2)\}\widetilde{x}]\|_\infty. \tag{11}$$

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 $\gamma = \gamma_n = 0.1/\log n$ and $c_0 = 1.1$. Below we summarize required technical conditions.

Condition PLAD. Assume that $\|\eta_0\|_0 = s \ge 1$, $E(\widetilde{x}_j^2) = 1$, $|\mathbb{E}_n(\widetilde{x}_j^2) - 1| \le 1/2$ for all $1 \le j \le p$ 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 $f_i(\widetilde{x}_i^{\mathrm{T}}\eta_0) \ge \underline{f} > 0$ is bounded away from zero.

Condition PLAD is implied by Condition 1 after a normalizing the variables so that $E(\widetilde{x}_j^2) = 1$. The assumption on the conditional density is standard in the quantile regression literature even with fixed p or p increasing slower than n (see Koenker, 2005; Belloni et al., 2011, respectively).

We present bounds on the population prediction norm of the ℓ_1-LAD estimator. The bounds depend on the restricted eigenvalue proposed in Bickel et al. (2009), defined by

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

where $\widetilde{T} = \mathrm{supp}(\eta_0)$ and $\Delta_{c_0} = \{\delta \in \mathbb{R}^{p+1} : \|\delta_{\widetilde{T}^c}\|_1 \le 3c_0'\|\delta_{\widetilde{T}}\|_1\}$ ($\widetilde{T}^c = \{1, \ldots, p+1\}\backslash\widetilde{T}$). The following lemma follows directly from the proof of Theorem 2 in Belloni & Chernozhukov (2011) applied to a single quantile index.

LEMMA 1 (ESTIMATION ERROR OF ℓ_1-LAD). *Under Condition PLAD and using* $\lambda = c_0 n\Lambda(1-\gamma \mid \widetilde{x}) \lesssim n\log\{(p\vee n)/\gamma\}$, *we have with probability at least* $1-\gamma-o(1)$,

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

provided that

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

Lemma 1 establishes the rate of convergence in the population prediction norm for the ℓ_1-LAD estimator in a parametric setting. The extra growth condition required for identification is mild. For instance for many designs of interest we have $\inf_{\delta\in\Delta_{c_0}}\|x^{\mathrm{T}}\delta\|_{P,2}^3/E(|\widetilde{x}^{\mathrm{T}}\delta|^3)$ bounded away from zero (Belloni & Chernozhukov, 2011). For designs with bounded regressors we have

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

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 $\widehat{T}$

$$\widetilde{\eta} \in \arg\min_{\eta}\left\{\mathbb{E}_n(|y-\widetilde{x}^{\mathrm{T}}\eta|) : \eta_j = 0 \text{ if } j\notin\widehat{T}\right\}. \tag{12}$$

The following result characterizes the performance of the estimator in (12); see Theorem 5 in Belloni & Chernozhukov (2011) for the proof.

LEMMA 2 (ESTIMATION ERROR OF POST-ℓ_1-LAD). *Suppose that* $\mathrm{supp}(\widehat{\eta}) \subseteq \widehat{T}$ *and let* $\widehat{s} = |\widehat{T}|$. *Then under the same conditions as in Lemma* 1,

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

provided that

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

Lemma 2 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-LAD) and the overall number of components $\widehat{s}$. Once again the extra growth condition required for identification is mild.

Remark 2. In Step 1 of Algorithm 2 we use ℓ_1-LAD 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 on rates for $\|x^{\mathrm{T}}(\widehat{\beta}-\beta_0)\|_{P,2}$ instead of $\|\widetilde{x}^{\mathrm{T}}\widehat{\delta}\|_{P,2}$. However, it follows that

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

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

$$|\widehat{\alpha} - \alpha_0| \leq \|\widehat{\delta}_{\widetilde{T}}\| \leq \|\widetilde{x}^{\mathrm{T}}\widehat{\delta}\|_{P,2}/\bar{\kappa}_{c_0},$$

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

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

3·3. *Heteroscedastic Lasso*

In this section we consider the equation (4) of the form

$$d_i = x_i^{\mathrm{T}}\theta_0 + v_i, \; E(v_i \mid x_i) = 0,$$

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

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n\{(d - x^{\mathrm{T}}\theta)^2\} + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1, \tag{13}$$

where λ and $\widehat{\Gamma}$ are the associated penalty level and loadings which are potentially data-driven. We rely on the results of Belloni et al. (2012) on the performance of Lasso and post-Lasso that allow for heteroscedasticity and non-Gaussianity. According to Belloni et al. (2012), we use the following options for the penalty level and the loadings:

$$\begin{aligned} &\text{initial} \quad \widehat{\gamma}_j = \sqrt{\mathbb{E}_n\{x_j^2(d-\bar{d})^2\}}, \; \lambda = 2c\sqrt{n}\Phi^{-1}\{1-\gamma/(2p)\}, \\ &\text{refined} \quad \widehat{\gamma}_j = \sqrt{\mathbb{E}_n(x_j^2\widehat{v}^2)}, \qquad\quad \lambda = 2c\sqrt{n}\Phi^{-1}\{1-\gamma/(2p)\}, \end{aligned} \tag{14}$$

for $1 \leq j \leq p$, where $c > 1$ is a fixed constant, $\gamma \in (1/n, 1/\log n)$, $\bar{d} = \mathbb{E}_n(d)$ 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 HL. (i) There exists $s = s_n \geq 1$ such that $\|\theta_0\|_0 \leq s$. (ii) $E(d^2) \leq C_1, \min_{1\leq j\leq p} E(x_j^2) \geq c_1$, $E(v^2 \mid x) \geq c_1$ almost surely, and $\max_{1\leq j\leq p} E(|x_j d|^2) \leq C_1$. (iii) $\max_{1\leq j\leq p}\{E(|x_j v|^3)\}^{1/3}\sqrt{\log(n\vee p)} = o(n^{1/6})$. (iv) With probability $1-o(1)$, $\max_{1\leq j\leq p}|\mathbb{E}_n(x_j^2 v^2) - E(x_j^2 v^2)| \vee \max_{1\leq j\leq p}|\mathbb{E}_n(x_j^2 d^2) - E(x_j^2 d^2)| = o(1)$ and $\max_{1\leq i\leq n}\|x_i\|_\infty^2 s\log(n\vee p) = o(n)$. (v) With probability $1-o(1)$, $c_1 \leq \phi^x_{\min}(\ell_n s) \leq \phi^x_{\max}(\ell_n s) \leq C_1$.

Condition HL (i) verifies Condition AS in Belloni et al. (2012), while Conditions HL (ii)-(iv) verify Condition RF in Belloni et al. (2012). Lemma 3 in Belloni et al. (2012) 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 Belloni et al. (2012) (applied to this setting) is bounded away from zero with probability $1-o(1)$; see Lemma 4.1 in Bickel et al. (2009).

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

LEMMA 3 (ESTIMATION ERROR OF LASSO). *Suppose that Condition HL is satisfied. Setting* $\lambda = 2c\sqrt{n}\Phi^{-1}\{1-\gamma/(2p)\}$ *for* $c > 1$*, and using the penalty loadings as in* (14)*, we have with probability* $1 - o(1)$*,*

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

Associated with Lasso we can define the Post-Lasso estimator as

$$\widetilde{\theta} \in \arg\min_{\theta} \left\{\mathbb{E}_n\{(d - x^{\mathrm{T}}\theta)^2\} : \theta_j = 0 \text{ if } \widehat{\theta}_j = 0\right\} \text{ and set } \widetilde{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}.$$

That is, the Post-Lasso estimator is simply the least squares estimator applied to the regressors selected by Lasso in (13). Sparsity properties of the Lasso estimator $\widehat{\theta}$ under estimated weights follows similarly to the standard Lasso analysis derived in Belloni et al. (2012). 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 4 (PROPERTIES OF LASSO AND POST-LASSO). *Suppose that Condition HL is satisfied. Consider the Lasso estimator with penalty level and loadings specified as in Lemma* 3*. 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

$$\|x^{\mathrm{T}}(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{\frac{s\log(p \vee n)}{n}}.$$

4. AUXILIARY TECHNICAL RESULTS

In this section we collect some auxiliary technical results.

LEMMA 5. *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 all* $1 \le i \le n$*, and there exists a constant* K_n *such that* $\max_{1\le i\le n}\|x_i\|_\infty \le K_n$ *almost surely. Then for every* $\tau \in (0, 1/8)$*, with probability at least* $1 - 8\tau$*,*

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

Proof of Lemma 5. The proof depends on the following maximal inequality derived in Belloni et al. (2014).

LEMMA 6. *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* $1 - 4\tau - 4\delta$*,*

$$\max_{1\le j\le p} |n^{-1/2}{\textstyle\sum_{i=1}^n}\{z_{ij} - E(z_{ij})\}| \le \left\{4\sqrt{2\log(2p/\delta)}\, Q(1-\tau)\right\} \vee 2\max_{1\le j\le p} \text{median of } |n^{-1/2}{\textstyle\sum_{i=1}^n}\{z_{ij} - E(z_{ij})\}|,$$

where $Q(u) = u$*-quantile of* $\max_{1\le j\le p}\sqrt{n^{-1}\sum_{i=1}^n z_{ij}^2}$*.*

Going back to the proof of Lemma 5, let $z_{ij} = \zeta_i^2 x_{ij}^2$. By Markov's inequality, we have

$$\text{median of } |n^{-1/2}\textstyle\sum_{i=1}^n \{z_{ij} - E(z_{ij})\}| \leq \sqrt{2n^{-1}\textstyle\sum_{i=1}^n E(z_{ij}^2)} \leq K_n^2 \sqrt{(2/n)\textstyle\sum_{i=1}^n E(\zeta_i^4)},$$

and

$$(1-\tau)\text{-quantile of } \max_{1\leq j\leq p} \sqrt{n^{-1}\textstyle\sum_{i=1}^n z_{ij}^2} \leq (1-\tau)\text{-quantile of } K_n^2\sqrt{n^{-1}\textstyle\sum_{i=1}^n \zeta_i^4}$$
$$\leq K_n^2 \sqrt{\textstyle\sum_{i=1}^n E(\zeta_i^4)/(n\tau)}.$$

Hence the conclusion of Lemma 5 follows from application of Lemma 6 with $\tau = \delta$. □

LEMMA 7. *Under Condition* 1, *there exists* $\ell_n' \to \infty$ *such that with probability* $1 - o(1)$,

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

Proof of Lemma 7. The lemma follows from application of Theorem 4.3 in Rudelson & Zhou (2013).

LEMMA 8. *Consider* p*-vectors* $\widehat{\beta}$ *and* β_0 *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*

$$\|\widehat{\beta}^{(m)} - \beta_0\|_1 \leq 2\|\widehat{\beta} - \beta_0\|_1$$
$$\|x^{\mathrm{T}}\{\widehat{\beta}^{(2m)} - \beta_0\}\|_{2,n} \leq \|x^{\mathrm{T}}(\widehat{\beta} - \beta_0)\|_{2,n} + \sqrt{\phi_{\max}^x(m)/m}\|\widehat{\beta} - \beta_0\|_1.$$

Proof of Lemma 8. 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

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

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, $\|\widehat{\beta}^{(km+m)} - \widehat{\beta}^{(km)}\| \leq \|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_1/\sqrt{m}$. Then

$$\|x^{\mathrm{T}}\{\widehat{\beta} - \widehat{\beta}^{(2m)}\}\|_{2,n} = \|x^{\mathrm{T}}\textstyle\sum_{k\geq 3}\{\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\}\|_{2,n} \leq \textstyle\sum_{k\geq 3}\|x^{\mathrm{T}}\{\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\}\|_{2,n}$$
$$\leq \sqrt{\phi_{\max}^x(m)}\textstyle\sum_{k\geq 3}\|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\| \leq \sqrt{\phi_{\max}^x(m)}\textstyle\sum_{k\geq 2}\|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_1/\sqrt{m}$$
$$= \sqrt{\phi_{\max}^x(m)}\|\widehat{\beta} - \widehat{\beta}^{(m)}\|_1/\sqrt{m} \leq \sqrt{\phi_{\max}^x(m)}\|\widehat{\beta} - \beta_0\|_1/\sqrt{m},$$

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

REFERENCES

ANDREWS, D. W. (1994). Empirical process methods in econometrics. *Handbook of Econometrics* **4**, 2247–2294.

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. & HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2430.

BELLONI, A. & CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics* **39**, 82–130.

BELLONI, A. & CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**, 521–547.

BELLONI, A., CHERNOZHUKOV, V. & FERNANDEZ-VAL, I. (2011). Conditional quantile processes based on series or many regressors. *arXiv:1105.6154* .

BELLONI, A., CHERNOZHUKOV, V. & HANSEN, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies* .

BELLONI, A., CHERNOZHUKOV, V. & KATO, K. (2013). Robust inference in high-dimensional approximately sparse quantile regression models. *arXiv:1312.7186* .

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

DE LA PEÑA, V. H., LAI, T. L. & SHAO, Q.-M. (2009). *Self-normalized Processes: Limit Theory and Statistical Applications*. Springer.

JING, B.-Y., SHAO, Q.-M. & WANG, Q. (2003). Self-normalized Cramer-type large deviations for independent random variables. *Annals of Probability* **31**, 2167–2215.

KOENKER, R. (2005). *Quantile Regression*. Cambridge University Press.

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

LEE, S. (2003). Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric theory* **19**, 1–31.

NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics, the Harold Cramer Volume*, U. Grenander, ed. New York: John Wiley and Sons, Inc.

NEYMAN, J. (1979). $C(\alpha)$ tests and their use. *Sankhya* **41**, 1–21.

RUDELSON, M. & ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59**, 3434–3447.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

WANG, L. (2013). L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis* **120**, 135–151.