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

Printed in Great Britain

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

BY A. BELLONI

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

AND V. CHERNOZHUKOV

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

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

1. INTRODUCTION

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

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

where d_i is the main regressor and coefficient α_0 is the main parameter of interest. The vector x_i denotes other high-dimensional regressors or controls. The regression error ϵ_i is independent

of d_i and x_i and has median zero, that is, $\mathrm{pr}(\epsilon_i \leq 0) = 1/2$. The distribution function of ϵ_i is denoted by F_ϵ and admits a density function f_ϵ such that $f_\epsilon(0) > 0$. The assumption motivates the use of the least absolute deviation or median regression, suitably adjusted for use in high-dimensional settings. The framework (1) is of interest in program evaluation, where d_i represents the treatment or policy variable known a priori and whose impact we would like to infer (Robinson, 1988; Liang et al., 2004; Imbens, 2004). We shall also discuss a generalization to the case where there are many parameters of interest, including the case where the identity of a regressor of interest is unknown a priori.

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

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

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

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

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

to perform inference for α_0.

This approach is justified if (2) achieves perfect model selection with probability approaching unity, so that the estimator (3) has the oracle property. However conditions for perfect selection are very restrictive in this model, and, in particular, require strong separation of non-zero coefficients away from zero. If these conditions do not hold, the estimator $\widetilde{\alpha}$ does not converge to α_0 at the $n^{-1/2}$ rate, uniformly with respect to the underlying model, and so the usual inference breaks down (Leeb & Pötscher, 2005). We shall demonstrate the breakdown of such naive inference in Monte Carlo experiments where non-zero coefficients in β_0 are not significantly separated from zero.

The breakdown of standard inference does not mean that the aforementioned procedures are not suitable for prediction. Indeed, the estimators (2) and (3) attain essentially optimal rates $\{(s \log p)/n\}^{1/2}$ of convergence for estimating the entire median regression function (Belloni & Chernozhukov, 2011; Wang, 2013). This property means that while these procedures will not deliver perfect model recovery, they will only make moderate selection mistakes, that is, they omit controls only if coefficients are local to zero.

In order to provide uniformly valid inference, we propose a method whose performance does not require perfect model selection, allowing potential moderate model selection mistakes. The latter feature is critical in achieving uniformity over a large class of data generating processes, similarly to the results for instrumental regression and mean regression studied in Zhang & Zhang (2014) and Belloni et al. (2012; 2013; 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 achieved by the naive estimators implies the failure of uniform validity of inference and their semiparametric inefficiency (Leeb & Pötscher, 2008).

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

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

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

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

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

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

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

so the estimator of α_0 will be unaffected by estimation of $x_i^{\mathrm{T}}\beta_0$ even if β_0 is estimated at a slower rate than $n^{-1/2}$, that is, the rate of $o(n^{-1/4})$ would suffice. This slow rate of estimation of the nuisance function permits the use of non-regular estimators of β_0, such as post-selection or regularized estimators that are not $n^{-1/2}$ consistent uniformly over the underlying model. The orthogonalization ideas can be traced back to Neyman (1959) and also play an important role in doubly robust estimation (Robins & Rotnitzky, 1995).

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

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

Step (ii) provides estimates $\widehat{v}_i$ of v_i in (4) as $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widehat{\theta}$ or $\widehat{v}_i = d_i - x_i^{\mathrm{T}}\widetilde{\theta}$ $(i = 1, \ldots, n)$. The first is based on the heteroscedastic lasso estimator $\widehat{\theta}$, a version of the lasso of Tibshirani (1996), designed to address non-Gaussian and heteroscedastic errors (Belloni et al., 2012),

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

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

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

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

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

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

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

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

in distribution, where $\sigma_n^2 = 1/\{4 f_\epsilon^2 E(v_i^2)\}$ with $f_\epsilon = f_\epsilon(0)$ is the semi-parametric efficiency bound for regular estimators of α_0. In the low-dimensional case, if $p^3 = o(n)$, the asymptotic behavior of our estimator coincides with that of the standard median regression without selection or penalization, as derived in He & Shao (2000), which is also semi-parametrically efficient in this case. However, the behaviors of our estimator and the standard median regression differ dramatically, otherwise, with the standard estimator even failing to be consistent when $p > n$. Of course, this improvement in the performance comes at the cost of assuming sparsity.

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

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

We estimate Ω by the plug-in method and J by Powell's (1986) method. Furthermore, we show that the score statistic $nL_n(\alpha)$ can be used for testing the null hypothesis $\alpha = \alpha_0$, and converges in distribution to a χ_1^2 variable under the null hypothesis, that is,

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

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

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

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

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

The second set of main results addresses a more general setting by allowing p_1-dimensional target parameters defined via Huber's Z-problems to be of interest, with dimension p_1 potentially much larger than the sample size n, and also allowing for approximately sparse models instead of exactly sparse models. This framework covers a wide variety of semi-parametric models, including those with smooth and non-smooth score functions. We provide sufficient conditions to derive a uniform Bahadur representation, and establish uniform asymptotic normality, using central limit theorems and bootstrap results of 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 previous results from the setting with $p_1 \ll n$ (Huber, 1973; Portnoy, 1984, 1985; He & Shao, 2000) to that with $p_1 \gg n$.

In what follows, the ℓ_2 and ℓ_1 norms are denoted by $\|\cdot\|$ and $\|\cdot\|_1$, respectively, and the ℓ_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. We use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Denote by $\Phi(\cdot)$ the distribution function of the standard

normal distribution. We assume that the quantities such as p, s, and hence $y_i, x_i, \beta_0, \theta_0, T$ and T_d are all dependent on the sample size n, and allow for the case where $p = p_n \to \infty$ and $s = s_n \to \infty$ as $n \to \infty$. We shall omit the dependence of these quantities on n when it does not cause confusion. For a class of measurable functions $\mathcal{F}$ on a measurable space, let $\mathrm{cn}(\epsilon, \mathcal{F}, \|\cdot\|_{Q,2})$ denote its ϵ-covering number with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$, where Q is a finitely discrete measure on the space, and let $\mathrm{ent}(\varepsilon, \mathcal{F}) = \log \sup_Q \mathrm{cn}(\varepsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2})$ denote the uniform entropy number where $F = \sup_{f\in\mathcal{F}} |f|$.

2. The Methods, Conditions, and Results

2·1. *The methods*

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

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

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

In order to perform ℓ_1-penalized median regression and lasso, one has to choose the penalty levels suitably. We record our penalty choices in the Supplementary Material. Algorithm 1 relies on the post-selection estimators that refit the non-zero coefficients without the penalty term to reduce the bias, while Algorithm 2 relies on the penalized estimators. In Step (ii), instead of the lasso or the post-lasso estimators, Dantzig selector (Candes & Tao, 2007) and Gauss-Dantzig estimators could be used. Step (iii) of both algorithms relies on instrumental median regression (9). Alternatively, in this step, we can use a one-step estimator $\check{\alpha}$ defined by

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

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

2·2. *Regularity conditions*

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

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

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

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

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

Condition 1 (i) imposes the setting discussed in the previous section with the zero conditional median of the error distribution. Condition 1 (ii) imposes moment conditions on the structural errors and regressors to ensure good model selection performance of lasso applied to equation (4). Condition 1 (iii) imposes sparsity of the high-dimensional vectors β_0 and θ_0. Condition 1 (iv) is a set of standard assumptions in median regression (Koenker, 2005) and in instrumental quantile regression. Condition 1 (v) restricts the sparsity index, namely $s^3 \log^3(p \vee n) = o(n)$ is required; this is analogous to the restriction $p^3 (\log p)^2 = o(n)$ made in He & Shao (2000) in the low-dimensional setting. The uniformly bounded regressors condition can be relaxed with minor modifications provided the bound holds with probability approaching unity. Most importantly, no assumptions on the separation from zero of the non-zero coefficients of θ_0 and β_0 are made. Condition 1 (vi) is quite plausible for many designs of interest. Conditions 1 (iv) and (v) imply the equivalence between the norms induced by the empirical and population Gram matrices over s-sparse vectors by Rudelson & Zhou (2013).

2·3. *Results*

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

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

Theorem 1 shows that Algorithms 1 and 2 produce estimators $\check{\alpha}$ that perform equally well, to the first order, with asymptotic variance equal to the semi-parametric efficiency bound; see the Supplemental Material for further discussion. Both algorithms rely on sparsity of $\widehat{\beta}$ and $\widehat{\theta}$. Sparsity of the latter follows immediately under sharp penalty choices for optimal rates. The sparsity for the former potentially requires a higher penalty level, as shown in Belloni & Chernozhukov (2011); alternatively, sparsity for the estimator in Step 1 can also be achieved by truncating the smallest components of $\widehat{\beta}$. The Supplemental Material shows that suitable truncation leads to the required sparsity while preserving the rate of convergence.

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

COROLLARY 1. *Let $\check{\alpha}$ 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),*

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

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

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

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

2·4. *Generalization to many target coefficients*

We consider the generalization to the previous model:

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

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

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

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

where α_j is the target coefficient,

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

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

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

include the standard Sobolev spaces as well as more general rearranged Sobolev spaces (Bickel et al., 2009; Belloni et al., 2014b). Here sparsity of estimators $\widehat{g}_j$ and $\widehat{m}_j$ means that they are formed by $O_P(s)$-sparse linear combinations chosen from p technical regressors generated from z_j, with coefficients estimated from the data. This framework is general; in particular it contains as a special case the traditional linear sieve/series framework for estimation of h_j, which uses a small number $s = o(n)$ of predetermined series functions as a dictionary.

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

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

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

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

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

3. INFERENCE ON MANY TARGET PARAMETERS IN Z-PROBLEMS

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

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

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

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

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

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

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

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

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

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

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

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

where we use the symbol ∂_t to abbreviate $\partial/\partial t$. It is important to construct the scores ψ_j to have property (18) or its generalization given in Remark 1 below. Generally, we can construct the scores ψ_j that obey such properties by projecting some initial non-orthogonal scores onto the orthogonal complement of the tangent space for the nuisance parameter (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 $b_n = \{\log(ep_1)\}^{1/2}$.

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

Condition 2 states rather mild assumptions for Z-estimation problems, in particular, allowing for non-smooth scores ψ_j such as those arising in median regression. They are analogous to assumptions imposed in the setting with $p = o(n)$, for example, in He & Shao (2000). The following condition uses a notion of pointwise measurable classes of functions (van der Vaart & Wellner, 1996, p.110).

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

has measurable envelope $F_j \geq \sup_{f \in \mathcal{F}_j} |f|$, such that $F = \max_{j=1,\ldots,p_1} F_j$ obeys $E\{F^q(w)\} \leq C_1$ for some $q \geq 4$; and (iv) the dimensions p_1, p, and s obey the growth conditions:

$$n^{-1/2}\{(s \log a_n)^{1/2} + n^{-1/2+1/q} s \log a_n\} \leq \rho_n, \;\; \rho_n^{\varsigma/2}(L_{2n} s \log a_n)^{1/2} + n^{1/2} L_{1n} \rho_n^2 \leq \delta_n b_n^{-1}.$$

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

Condition 3 (ii) implicitly requires ψ_j not to increase entropy too much; it holds, for example, when ψ_j is a monotone transformation, as in the case of median regression, or a Lipschitz transformation; see van der Vaart & Wellner (1996). The entropy bound is formulated in terms of the upper bound s on the sparsity of the estimators and p the dimension of the overall approximating model appearing via a_n. In principle our main result below applies to non-sparse estimators as well, as long as the entropy bound specified in Condition 3 (ii) holds, with index (s, p) interpreted as measures of effective complexity of the relevant function classes.

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

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

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

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

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

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

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

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

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

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

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

Another useful implication is the high-dimensional central limit theorem uniformly over rectangles in $\mathbb{R}^{p_1}$, provided that $(\log p_1)^7 = o(n)$, which follows from Corollary 2.1 in Chernozhukov et al. (2013). Let $\mathcal{N} = (\mathcal{N}_j)_{j=1}^p$ be a normal random vector in $\mathbb{R}^{p_1}$ with mean zero and covariance matrix $[E\{\phi_j(w)\phi_{j'}(w)\}]_{j,j'=1}^{p_1}$. Let $\mathcal{R}$ be a collection of rectangles R in $\mathbb{R}^{p_1}$ of the

form

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

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

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

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

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

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

This result leads to simultaneous confidence bands for $(\alpha_j)_{j=1}^{p_1}$ that are valid uniformly in $P \in \mathcal{P}_n$. Moreover, Corollary 3 is immediately useful for testing multiple hypotheses about $(\alpha_j)_{j=1}^{p_1}$ via the step-down methods of Romano & Wolf (2005) which control the family-wise error rate; 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, since its covariance matrix is unknown, but it can be approximated by the Gaussian multiplier bootstrap, which generates a vector

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

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

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

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

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

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

Remark 1. The proof of Theorem 2 shows that the orthogonality condition (18) can be replaced by a more general orthogonality condition:

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

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

4. MONTE CARLO EXPERIMENTS

We consider the regression model

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

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

The first equation in (21) is a sparse model. However, unless c_y is very large, the decay of the components of θ_0 rules out the typical assumption that the coefficients of important regressors are well separated from zero. Thus we anticipate that the standard post-selection inference procedure, discussed around (3), would work poorly in the simulations. In contrast, from the prior theoretical arguments, we anticipate that our instrumental median estimator would work well.

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

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

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

ACKNOWLEDGMENTS

This paper was first presented in 8th World Congress in Probability and Statistics in August 2012. We would like to thank the editor, an associate editor, and anonymous referees for their careful review. We are also grateful to Sara van de Geer, Xuming He, Richard Nickl, Roger Koenker, Vladimir Koltchinskii, Enno Mammen, Steve Portnoy, Philippe Rigollet, Richard Samworth, and Bin Yu for useful comments and discussions. Research support from the National Science Foundation and the Japan Society for the Promotion of Science is gratefully acknowledged.

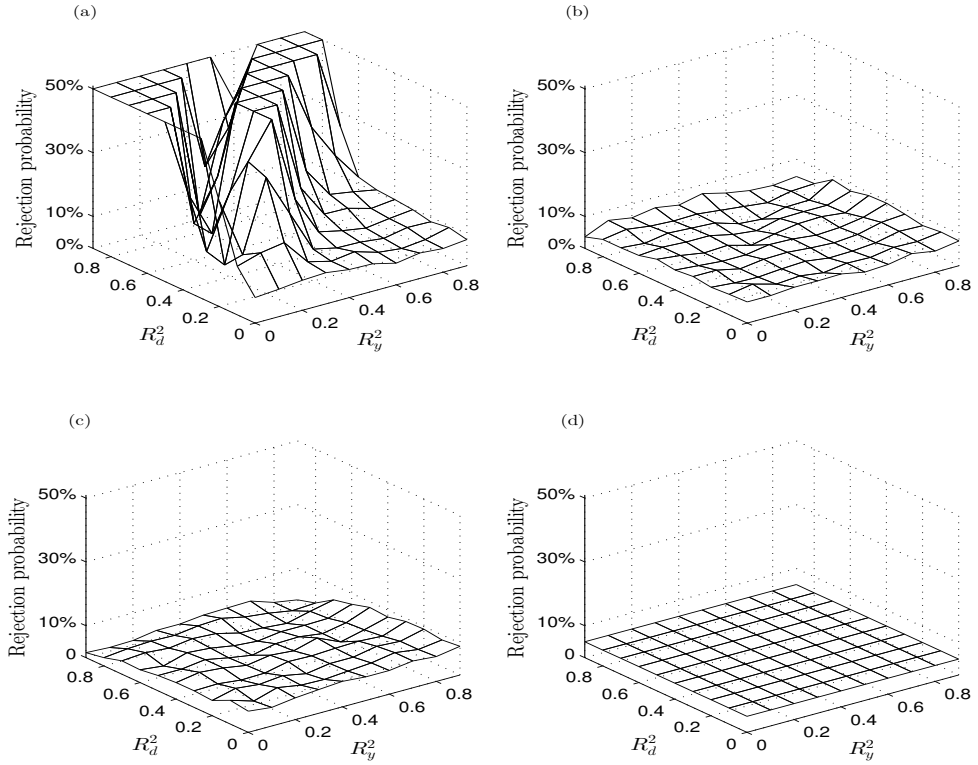

Fig. 1. The empirical false rejection probabilities of the nominal 5% level tests based on: (a) the standard post-model selection procedure based on $\widetilde{\alpha}$, (b) the proposed post-model selection procedure based on $\check{\alpha}$, (c) the score statistic L_n, and (d) an ideal procedure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

SUPPLEMENTARY MATERIAL

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

A. PROOF OF THEOREM 2

A·1. *A maximal inequality*

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

LEMMA A1. *Let $w, w_1, \ldots, w_n$ be independent and identically distributed random variables taking values in a measurable space, and let $\mathcal{F}$ be a pointwise measurable class of functions on that space. Suppose that there is a measurable envelope $F \geq \sup_{f\in\mathcal{F}} |f|$ such that $E\{F^q(w)\} < \infty$ for some $q \geq 2$. Consider the empirical process indexed by $\mathcal{F}$: $G_n(f) = n^{-1/2}\sum_{i=1}^n [f(w_i) - E\{f(w)\}]$, $f \in \mathcal{F}$. Let $\sigma > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} E\{f^2(w)\} \leq \sigma^2 \leq E\{F^2(w)\}$. Moreover, suppose that*

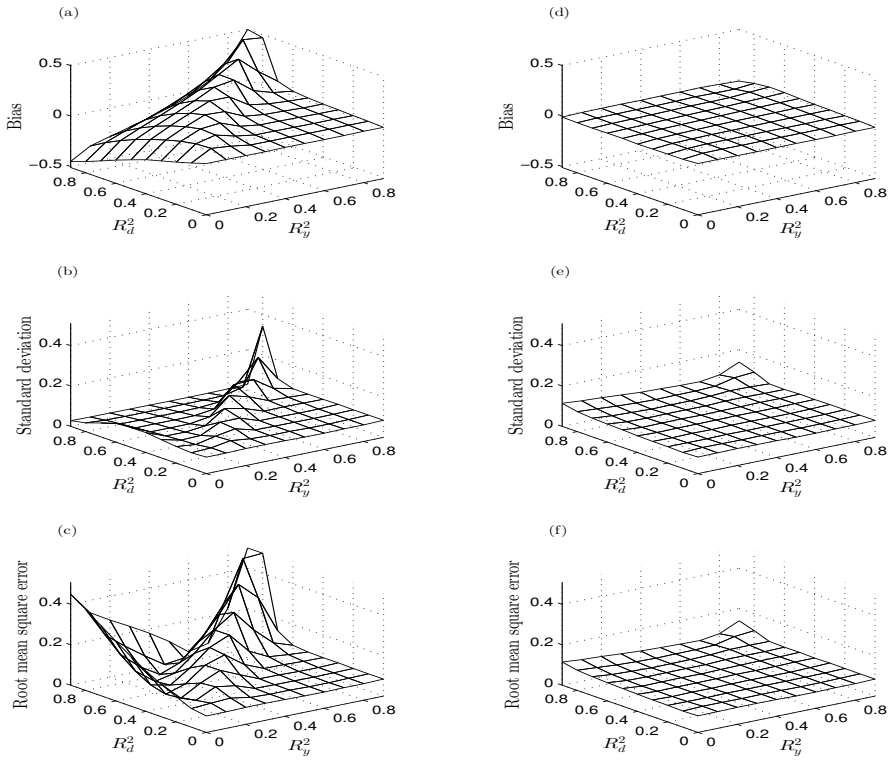

Fig. 2. Mean bias (top row), standard deviation (middle row), root mean square (bottom row) of the standard post-model selection estimator $\widetilde{\alpha}$ (panels (a)-(c)), and of the proposed post-model selection estimator $\check{\alpha}$ (panels (d)-(f)).

there exist constants $A \geq e$ and $s \geq 1$ such that $\mathrm{ent}(\varepsilon, \mathcal{F}) \leq s \log(A/\varepsilon)$ *for all* $0 < \varepsilon \leq 1$. *Then*

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

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

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

where K_q is a constant that depends only on q.

Proof. The first and second inequalities follow from Corollary 5.1 and Theorem 5.1 in Chernozhukov et al. (2014) applied with $\alpha = 1$, using that $[E\{\max_{i=1,\ldots,n} F^2(w_i)\}]^{1/2} \leq [E\{\max_{i=1,\ldots,n} F^q(w_i)\}]^{1/q} \leq n^{1/q}[E\{F^q(w)\}]^{1/q}$. □

A·2. *Proof of Theorem* 2

It suffices to prove the theorem under any sequence $P = P_n \in \mathcal{P}_n$. We shall suppress the dependence of P on n in the proof. In this proof, let C denote a generic positive constant that may differ in each appearance, but that does not depend on the sequence $P \in \mathcal{P}_n$, n, or $j = 1, \ldots, p_1$. Recall that the sequence $\rho_n \downarrow 0$ satisfies the growth conditions in Condition 3 (iv). We divide the proof into three steps. Below we use the following notation: for any given function $g : \mathcal{W} \to \mathbb{R}$, $G_n(g) = n^{-1/2} \sum_{i=1}^n [g(w_i) - E\{g(w)\}]$.

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

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

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

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

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

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

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

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

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 $m = 1, \ldots, M$ whenever $\widetilde{h} = (\widetilde{h}_m)_{m=1}^M \in \mathcal{H}_j$. Hence applying Lemma 1 with $t = \log n$, we conclude that, with probability $1 - o(1)$,

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

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

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

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

The third term is zero because of the orthogonality condition (18). Condition 2 (ii) guarantees that the expectation and derivative can be interchanged for the second term, that is, $E\left(\partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j]\right) = \partial_\alpha E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = \Gamma_j$. Moreover, by the same condition, each of the last three terms is bounded by $CL_{1n}\rho_n^2 = o(n^{-1/2} b_n^{-1})$, uniformly in $j = 1, \ldots, p_1$. Therefore, with probability $1 - o(1)$, $II_j = \Gamma_j(\widetilde{\alpha}_j - \alpha_j) + o(n^{-1/2} b_n^{-1})$, uniformly in $j = 1, \ldots, p_1$. Combining the previous bound on III_j with these bounds leads to the desired assertion.

Step 2. We wish to show that with probability $1-o(1)$, $\inf_{\alpha\in\widehat{\mathcal{A}}_j}|E_n[\psi_j\{w,\alpha,\widehat{h}_j(z_j)\}]| = o(n^{-1/2}b_n^{-1})$, uniformly in $j=1,\dots,p_1$. Define $\alpha_j^* = \alpha_j - \Gamma_j^{-1}E_n[\psi_j\{w,\alpha_j,h_j(z_j)\}]$ $(j=1,\dots,p_1)$. Then we have $\max_{j=1,\dots,p_1}|\alpha_j^*-\alpha_j| \le C\max_{j=1,\dots,p_1}|E_n[\psi_j\{w,\alpha_j,h_j(z_j)\}]|$. Consider the class of functions $\mathcal{F}' = \{w\mapsto\psi_j\{w,\alpha_j,h_j(z_j)\} : j=1,\dots,p_1\}$, which has F as an envelope. Since this class is finite with cardinality p_1, we have $\mathrm{ent}(\varepsilon,\mathcal{F}')\le\log(p_1/\varepsilon)$. 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 $1-o(1)$,

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

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 $1-o(1)$,

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

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

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

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

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

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

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 $j=1,\dots,p_1$, 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 $m=1,\dots,M$ 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 $j=1,\dots,p_1$.

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

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

uniformly in $j=1,\dots,p_1$. Moreover, by Step 2, with probability $1-o(1)$, the left side is $o(n^{-1/2}b_n^{-1})$ uniformly in $j=1,\dots,p_1$. 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. *Ann. Statist.* **39**, 82–130.

BELLONI, A., CHERNOZHUKOV, V. & HANSEN, C. (2013). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics: The 2010 World Congress of the Econometric Society* **3**, 245–295.

BELLONI, A., CHERNOZHUKOV, V. & HANSEN, C. (2014a). Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.* **81**, 608–650.

BELLONI, A., CHERNOZHUKOV, V. & WANG, L. (2014b). Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.* **42**, 757–788.

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

CANDES, E. & TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

CHERNOZHUKOV, V., CHETVERIKOV, D. & KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41**, 2786–2819.

CHERNOZHUKOV, V., CHETVERIKOV, D. & KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42**, 1564–1597.

CHERNOZHUKOV, V. & HANSEN, C. (2008). Instrumental variable quantile regression: A robust inference approach. *J. Econometrics* **142**, 379–398.

HE, X. & SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73**, 120–135.

HUBER, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86**, 4–29.

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

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: 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. *J. Econometrics* **142**, 201–211.

LIANG, H., WANG, S., ROBINS, J. M. & CARROLL, R. J. (2004). Estimation in partially linear models with missing covariates. *J. Amer. Statist. Assoc.* **99**, 357–367.

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.

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

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

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

ROBINS, J. M. & ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90**, 122–129.

ROBINSON, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* **56**, 931–954.

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 Trans. Inform. Theory* **59**, 3434–3447.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B* **58**, 267–288.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer-Verlag.

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

ZHANG, C.-H. & ZHANG, S. S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B* **76**, 217–242.