cemmap
centre for microdata methods and practice

Valid post-selection inference in high-dimensional approximately sparse quantile regression models

Alexandre Belloni
Victor Chernozhukov
Kengo Kato

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP53/14

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

VALID POST-SELECTION INFERENCE IN HIGH-DIMENSIONAL APPROXIMATELY SPARSE QUANTILE REGRESSION MODELS

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

Abstract. This work proposes new inference methods for the estimation of a regression coefficient of interest in quantile regression models. We consider high-dimensional models where the number of regressors potentially exceeds the sample size but a subset of them suffice to construct a reasonable approximation of the unknown quantile regression function in the model. The proposed methods are protected against moderate model selection mistakes, which are often inevitable in the approximately sparse model considered here. The methods construct (implicitly or explicitly) an optimal instrument as a residual from a density-weighted projection of the regressor of interest on other regressors. Under regularity conditions, the proposed estimators of the quantile regression coefficient are asymptotically root-n normal, with variance equal to the semi-parametric efficiency bound of the partially linear quantile regression model. In addition, the performance of the technique is illustrated through Monte-carlo experiments and an empirical example, dealing with risk factors in childhood malnutrition. The numerical results confirm the theoretical findings that the proposed methods should outperform the naive post-model selection methods in non-parametric settings. Moreover, the empirical results demonstrate soundness of the proposed methods.

1. Introduction

Many applications of interest requires the measurement of the distributional impact of a policy (or treatment) on the relevant outcome variable. Quantile treatment effects have emerged as an important concepts for measuring such distributional impact (see, e.g., [22]). In this work we focus on the quantile treatment effect α_τ of a policy/treatment d of an outcome of interest y in the partially linear model:

$$\tau - \text{quantile}(y \mid z, d) = d\alpha_\tau + g_\tau(z).$$

Here α_τ is the quantile treatment effect ([29, 22]), and g_τ is the confounding effects of the other covariates or controls z. To approximate g_τ we rely on linear combinations of p-dimensional vector of technical regressors, $x = P(z)$, where we allow for the dimension p to be potentially bigger than the sample size n to achieve an accurate approximation for g_τ. This brings forth the need to perform model selection or regularization.

We propose methods to construct estimates and confidence regions for the coefficient of interest α_τ, based upon robust post-selection procedures. We establish the (uniform) validity of the proposed methods in a non-parametric setting. Model selection in those settings (generically) leads to a (moderate) misspecification of the selected model and traditional arguments based on perfect model selection do not

Date: First version: May 2012, this version December 30, 2014.

apply. Therefore the proposed methods are developed to be robust to (moderate) model selection mistakes. The proposed methods achieve the asymptotic semi-parametric efficiency bound for the partially linear quantile regression model. To do so the conditional densities should be used as weights in the second step of the method. Typically such density function is unknown and needs to be estimated which leads to high dimensional model selection problems with estimated data. [1]

The proposed methods proceed in three steps. The first step aims to construct an estimate of the control function g_τ. This can be achieved via ℓ_1-penalized quantile regression estimator [3, 19, 39] or quantile regression post-selection based on ℓ_1-penalized quantile regression [3]. The second step attempts to properly partial out the confounding factors z from the treatment. The heteroscedasticity in the model requires us to consider a density-weighted equation, whose estimation is carried out by the heteroscedastic post-Lasso [34, 2]. The third step combines the estimates above to construct an estimate of α_τ which is robust to the non-regular estimation in the previous steps. The fact that the estimators in the first two steps are non-regular is a generic feature of our problem. We propose to implement this last step via instrumental quantile regression [15] or by a density-weighted quantile regression with all the variables selected in the previous steps, with the latter step reminiscent of the "post-double selection" method proposed in [6, 7]. Explicitly or implicitly the third step estimates α_τ by minimizing the Neyman-type score statistic. We mostly focus on selection as a means of regularization, but certainly other regularization (e.g. the use of ℓ_1-penalized fits per se) is possible, thought performs less well than the methods we focus on.

Our paper contributes to the new literature on inference (as opposed to estimation) in the high-dimensional sparse models. Several recent papers study the problem of constructing confidence regions after model selection allowing $p \gg n$. In the case of linear mean regression, [6] proposed a double selection inference in a parametric with homoscedastic Gaussian errors, [7] studies the double selection procedure in a non-parametric setting with heteroscedastic errors, [40] and [36] proposed estimators based on ℓ_1-penalized estimators based on "1-step" correction in parametric models. Going beyond mean models, [36] also provides high level conditions for the one-step estimator applied to smooth generalized linear problems, [8] analyzes confidence regions for a parametric homoscedastic LAD regression under primitive conditions based on the instrumental LAD regression, and [10] provides two post-selection procedures to build confidence regions for the logistic regression. None of the aforementioned papers deal with the problem of the present paper.

Some of the papers above explicitly (or implicitly) aim to achieve an important uniformity guarantees with respect to the (unknown) values of the parameters. These uniform properties translate into more reliable finite sample performance of these inference procedures because they are robust with respect to (unavoidable) model selection mistakes. There is now substantial theoretical and empirical evidence on the potential poor finite sample performance of estimators that rely on perfect model selection to build confidence regions when applied to models without separation from zero of the coefficients (i.e. small coefficients). Most of the criticism of these procedures are consequence of negative results established

[1] We also discuss alternative estimators that avoid the use of model selection procedures with estimated data. Those can be valid under weaker conditions, but they are not semi-parametric efficient, except for some special (homoscedastic) cases.

in [26], [28] and the references therein. This work contributes to this literature by proposing methods that will deliver confidence regions that also have uniformity guarantees for (heteroscedastic) quantile regression models allowing $p \gg n$. Although related in spirit with our previous work, [7, 8, 10], new tools and major departures are required to accommodate the non-differentiability of the loss function, heteroscedsaticity of the data, and the non-parametric setting.

Finally, in the process of establishing the main results we also contribute to the literature of high-dimensional estimation. An intermediary step of the method required the estimation of a weighted least squares version of Lasso in which weights are estimated. Finite sample bounds of Lasso for the prediction rate are established to this new case. Finite sample bounds for the prediction norm on the estimation error of ℓ_1-penalized quantile regression in nonparametric models extending results on [3, 19, 39]. We further developed results on instrumental quantile regression problems in which we allow for the dimension to increase and estimated instruments.

Notation. In what follows, we work with triangular array data $\{(\omega_{i,n}, i = 1, ..., n), n = 1, 2, 3, ...\}$ defined on probability space $(\Omega, \mathcal{S}, \mathrm{P}_n)$, where $\mathrm{P} = \mathrm{P}_n$ can change with n. Each $\omega_{i,n} = (y'_{i,n}, z'_{i,n}, d'_{i,n})'$ is a vector with components defined below, and these vectors are i.n.i.d. – independent across i, but not necessarily identically distributed. Thus, all parameters that characterize the distribution of $\{\omega_{i,n}, i = 1, ..., n\}$ are implicitly indexed by P_n and thus by n. We omit the dependence from the notation in what follows for notational simplicity. We use array asymptotics to better capture some finite-sample phenomena and to insure the robustness of conclusions with respect to perturbations of the data-generating process P along various sequences. We use $\mathbb{E}_n$ to abbreviate the notation $n^{-1}\sum_{i=1}^n$ and the following empirical process notation, $\mathbb{E}_n[f] := \mathbb{E}_n[f(\omega_i)] := \sum_{i=1}^n f(\omega_i)/n$, and $\mathbb{G}_n(f) := \sum_{i=1}^n (f(\omega_i) - \mathrm{E}[f(\omega_i)])/\sqrt{n}$. Since we want to deal with i.n.i.d. data, we also introduce the average expectation operator: $\bar{\mathrm{E}}[f] := \mathrm{E}\mathbb{E}_n[f] = \mathrm{E}\mathbb{E}_n[f(\omega_i)] = \sum_{i=1}^n \mathrm{E}[f(\omega_i)]/n$. The l_2-norm is denoted by $\|\cdot\|$, and the l_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. We use $\|\cdot\|_\infty$ to denote the maximal 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 in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We let $\delta^{(k)}$ be a vector with k non-zero components corresponding to k of the largest components of δ in absolute value. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E, we say that E wp $\to 1$ when E occurs with probability approaching one as n grows. Given a p-vector b, we denote support$(b) = \{j \in \{1, ..., p\} : b_j \neq 0\}$. We also use $\rho_\tau(t) = t(\tau - 1\{t \leqslant 0\})$ and $\varphi_\tau(t_1, t_2) = (\tau - 1\{t_1 \leqslant t_2\})$.

2. Setting and Methods

For a quantile index $\tau \in (0, 1)$, we consider the following partially linear conditional quantile model

$$y_i = d_i\alpha_\tau + g_\tau(z_i) + \epsilon_i, \quad \tau - \text{quantile}(\epsilon_i \mid d_i, z_i) = 0, \quad i = 1, \ldots, n, \tag{2.1}$$

where y_i is the outcome variable, d_i is the policy/treatment variable, and confounding factors are represented by the variables z_i which impacts the equation through an unknown function g_τ. The main

parameter of interest is α_τ, which is the quantile treatment effect, which describes the impact of the treatment on the conditional quantiles. We assume that the disturbance term ϵ_i in (2.1) has a positive and finite conditional density at 0,

$$f_i = f_{\epsilon_i}(0 \mid d_i, z_i). \tag{2.2}$$

We shall use a large number p of technical controls $x_i = P(z_i)$ to achieve an accurate approximation to the functions g_τ in (2.1) which take the form:

$$g_\tau(z_i) = x_i'\beta_\tau + r_{g\tau i}, \quad i = 1, \ldots, n, \tag{2.3}$$

where $r_{g\tau i}$ denotes an approximation error.

In order to perform robust inference with respect to model selection mistakes, we also consider an instrumental variable $\iota_{0i} = \iota_0(d_i, z_i)$ with the properties:

$$\bar{\mathrm{E}}[(1\{y_i \leqslant d_i\alpha_\tau + g_\tau(z_i)\} - \tau)\iota_{0i}] = 0, \tag{2.4}$$

$$\tfrac{\partial}{\partial\alpha}\bar{\mathrm{E}}[(1\{y_i \leqslant d_i\alpha + g_\tau(z_i)\} - \tau)\iota_{0i}]\big|_{\alpha=\alpha_\tau} = \bar{\mathrm{E}}[f_i\iota_{0i}d_i] \neq 0, \tag{2.5}$$

and

$$\frac{\partial}{\partial\delta}\bar{\mathrm{E}}[(1\{y_i \leqslant d_i\alpha + g_\tau(z_i) + \delta' x_i\} - \tau)\iota_{0i}]\bigg|_{\delta=0} = \bar{\mathrm{E}}[f_i\iota_{0i}x_i] = 0. \tag{2.6}$$

The relations (2.4)-(2.5) provide the estimating equation as well as the identification condition for α_τ. Relation (2.6) states that the estimating equation should be immune/insensitive to local perturbations of the nuisance function g_τ in the directions spanned by x_i. This orthogonality property is the critical ingredient in guaranteeing robustness of procedures, proposed below, against the preliminary "crude" estimation of the nuisance function g_τ. In particular, this ingredient delivers robustness to moderate model selection mistakes that accrue when post-selection estimators of g_τ are used.

The (optimal) instrument satisfying (2.4) and (2.6) can be defined as the residual v_i in the following decomposition for the regressor of interest d_i weighted by the conditional density function, namely

$$f_i d_i = f_i x_i'\theta_{0\tau} + v_i, \quad \mathrm{E}[f_i v_i x_i] = 0, \quad i = 1, \ldots, n, \tag{2.7}$$

and, thus, the (optimal) instrument is

$$\iota_{0i} = v_i = f_i d_i - f_i x_i'\theta_{0\tau}. \tag{2.8}$$

We should point that we can construct other (non-optimal) instruments satisfying (2.4) by using different weights $\tilde{f}_i$ instead of f_i in the equation (2.7) and setting $\tilde{\iota}_{0i} = \tilde{v}_i(\tilde{f}_i/f_i)$ where $\tilde{v}_i$ is the new residual corresponding to $\tilde{f}_i$. It turns out that the choice $\tilde{f}_i = f_i$ minimizes the asymptotic variance of the estimator of $\widehat{\alpha}$ based upon the empirical analog of (2.4), among all the instruments satisfying (2.5) and (2.6). We note that the problem of constructing optimal estimating equation (via optimal instruments) is equivalent to constructing the optimal score for the parameter α_τ.

We assume that β_τ and $\theta_{0\tau}$ are approximately sparse, namely it is possible to choose the parameters β_τ and θ_τ such that:

$$\|\theta_\tau\|_0 \leqslant s, \quad \|\beta_\tau\|_0 \leqslant s, \quad \bar{\mathrm{E}}[r_{\theta\tau i}^2] \lesssim s/n \text{ and } \bar{\mathrm{E}}[r_{g\tau i}^2] \lesssim s/n \tag{2.9}$$

where $r_{\theta\tau i} = x_i'\theta_{0\tau} - x_i'\theta_\tau$ and $r_{g\tau i} = g(z_i) - x_i'\beta_\tau$. The latter equation requires that it is possible to choose the sparsity index s so that the mean squared approximation error is of no larger order than the variance of the oracle estimator for estimating the coefficients in the approximation. (See [14] for a detailed discussion of this notion of approximately sparsity.)

Comment 2.1 (Handling Approximately Sparse Models)**.** In order to handle approximately sparse models to represent g_τ in (2.3), we will assume that the approximation errors $r_{g\tau}$ are nearly orthogonal with respect to fv, namely

$$\bar{\mathrm{E}}[f_i v_i r_{g\tau i}] = o(n^{-1/2}). \tag{2.10}$$

Condition (2.10) is automatically satisfied if the model is exactly sparse $r_{g\tau} = 0$ or if the orthogonality condition in (2.7) can be strengthen to $\mathrm{E}[f_i v_i \mid z_i] = 0$. Both (more stringent) assumptions have been used in the literature. However, (2.10) can be satisfied in many other cases by the orthogonality of v_i with respect to x_i in (2.7) in high-dimensional settings. We defer to Section 5.3 for a detailed discussion on how the high-dimensionality can yield (2.10) when g_τ belongs to a well behaved class of functions like a Sobolev ball.

2.1. **Known Conditional Density Function.** In this subsection we consider the case of known conditional density function f_i. This case is of theoretical value since it allows to abstract away from estimating the conditional density function f_i and focus on the principal features of the problem. Moreover, under homoscedasticity, when $f_i = f$ for all i, the unknown constant f will cancel in the definition of the estimators proposed below and the results are also of practical interest for that case. In what follows, we use the normalization $\mathbb{E}_n[x_{ij}^2] = 1$, $j = 1, \ldots, p$, to define the algorithms and collect the recommended choice of tuning parameters in Remark 2.3 below. Recall that for a vector β, $\beta^{(2s)}$ will truncate to zero all components of β except the $2s$ largest components in absolute value.

We will consider two procedures in detail. They are based on ℓ_1-penalized quantile regression and ℓ_1-penalized weighted least squares. The first procedure (Algorithm 1) is based on the explicit construction of the optimal instruments (2.8) and the use of instrumental quantile regression.

Algorithm 1 (Instrumental Quantile Regression based on Optimal Instrument)

(1) Run Post-ℓ_1-quantile regression of y_i on d_i and x_i; keep fitted value $x_i'\widetilde{\beta}_\tau$,

$$(\widehat{\alpha}_\tau, \widehat{\beta}_\tau) \in \arg\min_{\alpha,\beta} \mathbb{E}_n[\rho_\tau(y_i - d_i\alpha - x_i'\beta)] + (\lambda_\tau/n)\|\beta\|_1 + (\lambda_\tau/n)\{\mathbb{E}_n[d_i^2]\}^{1/2}|\alpha|$$
$$(\widetilde{\alpha}_\tau, \widetilde{\beta}_\tau) \in \arg\min_{\alpha,\beta} \mathbb{E}_n[\rho_\tau(y_i - d_i\alpha - x_i'\beta)] \ : \ \mathrm{support}(\beta) \subseteq \mathrm{support}(\widehat{\beta}_\tau^{(2s)}).$$

(2) Run Post-Lasso of $f_i d_i$ on $f_i x_i$; keep the residual $\widetilde{v}_i := f_i(d_i - x_i'\widetilde{\theta}_\tau)$,

$$\widehat{\theta}_\tau \in \arg\min_\theta \mathbb{E}_n[f_i^2(d_i - x_i'\theta)^2] + (\lambda/n)\|\widehat{\Gamma}_\tau\theta\|_1$$
$$\widetilde{\theta}_\tau \in \arg\min_\theta \mathbb{E}_n[f_i^2(d_i - x_i'\theta)^2] \ : \ \mathrm{support}(\theta) \subseteq \mathrm{support}(\widehat{\theta}_\tau).$$

(3) Run Instrumental Quantile Regression of $y_i - x_i'\widetilde{\beta}_\tau$ on d_i using $\widetilde{v}_i$ as the instrument for d_i,

$$\check{\alpha}_\tau \in \arg\min_{\alpha\in\mathcal{A}_\tau} L_n(\alpha), \text{ where } L_n(\alpha) := \frac{\{\mathbb{E}_n[(1\{y_i \leqslant d_i\alpha + x_i'\widetilde{\beta}_\tau\} - \tau)\widetilde{v}_i]\}^2}{\mathbb{E}_n[(1\{y_i \leqslant d_i\alpha + x_i'\widetilde{\beta}_\tau\} - \tau)^2\widetilde{v}_i^2]}, \text{ and set } \check{\beta}_\tau = \widetilde{\beta}_\tau.$$

Comment 2.2. In Algorithm 1 we can also work with the corresponding ℓ_1-penalized estimators in Steps 1 and 2 instead of the post-selection estimators, though we found that the latter work significantly better in computational experiments. □

The second procedure (Algorithm 2) creates the optimal instruments implicitly by using a weighted quantile regression based on double selection.

Algorithm 2 (Weighted Quantile Regression based on Double Selection)

(1) Run ℓ_1-quantile regression of y_i on d_i and x_i,

$$(\widehat{\alpha}_\tau, \widehat{\beta}_\tau) \in \arg\min_{\alpha,\beta} \mathbb{E}_n[\rho_\tau(y_i - d_i\alpha - x_i'\beta)] + (\lambda_\tau/n)\|\beta\|_1 + (\lambda_\tau/n)\{\mathbb{E}_n[d_i^2]\}^{1/2}|\alpha|$$

(2) Run Lasso of $f_i d_i$ on $f_i x_i$,

$$\widehat{\theta}_\tau \in \arg\min_\theta \mathbb{E}_n[f_i^2(d_i - x_i'\theta)^2] + (\lambda/n)\|\widehat{\Gamma}_\tau\theta\|_1$$

(3) Run quantile regression of $f_i y_i$ on $f_i d_i$ and $\{f_i x_{ij}, j \in \text{support}(\widehat{\beta}_\tau^{(2s)}) \cup \text{support}(\widehat{\theta}_\tau)\}$,

$$(\check{\alpha}_\tau, \check{\beta}_\tau) \in \arg\min_{\alpha,\beta} \mathbb{E}_n[f_i\rho_\tau(y_i - d_i\alpha - x_i'\beta)] : \text{support}(\beta) \subseteq \text{support}(\widehat{\beta}_\tau^{(2s)}) \cup \text{support}(\widehat{\theta}_\tau),$$

and set $L_n(\alpha) := \{\mathbb{E}_n[(1\{y_i \leqslant d_i\alpha + x_i'\check{\beta}_\tau\} - \tau)\widetilde{v}_i]\}^2 / \mathbb{E}_n[(1\{y_i \leqslant d_i\alpha + x_i'\check{\beta}_\tau\} - \tau)^2\widetilde{v}_i^2]$, where $\widetilde{v}_i = f_i(d_i - x_i'\widetilde{\theta}_\tau)$, and $\widetilde{\theta}_\tau$ is the post-Lasso estimator associated with $\widehat{\theta}_\tau$.

Comment 2.3 (Choice of Parameters)**.** We normalize the regressors so that $\mathbb{E}_n[x_{ij}^2] = 1$ throughout the paper. For $\gamma = 0.05/\{n \vee p\log n\}$, we set the penalty levels as

$$\lambda := 1.1\sqrt{n}2\Phi^{-1}(1-\gamma), \quad \text{and} \quad \lambda_\tau := 1.1\sqrt{n\tau(1-\tau)}\Phi^{-1}(1-\gamma). \tag{2.11}$$

The penalty loading $\widehat{\Gamma}_\tau = \text{diag}[\widehat{\Gamma}_{\tau jj}, j = 1, ..., p]$ is a diagonal matrix defined by the the following procedure: (1) Compute the Post Lasso estimator $\widetilde{\theta}_\tau^0$ based on λ and initial values $\widehat{\Gamma}_{\tau jj} = \max_{i \leqslant n} f_i\{\mathbb{E}_n[x_{ij}^2 f_i^2 d_i^2]\}^{1/2}$. (2) Compute the residuals $\widehat{v}_i = f_i(d_i - x_i'\widetilde{\theta}_\tau^0)$ and update

$$\widehat{\Gamma}_{\tau jj} = \sqrt{\mathbb{E}_n[f_i^2 x_{ij}^2 \widehat{v}_i^2]}, \ j = 1, \ldots, p. \tag{2.12}$$

In Algorithm 1 we have used the following parameter space for the computations:

$$\mathcal{A}_\tau = \{\alpha \in \mathbb{R} : |\alpha - \widetilde{\alpha}_\tau| \leqslant 10\{\mathbb{E}_n[d_i^2]\}^{-1/2}/\log n\}. \tag{2.13}$$

Typically s is unknown and to implement the algorithm we recommend setting the truncation parameter to $\frac{10}{\log n}\left\{\log n + \frac{n^{1/3}}{\log(p\vee n)} \wedge \frac{n^{1/2}\log^{-3/2}(p\vee n)}{\max_{i\leqslant n}\|x_i\|_\infty}\right\}$. Note that if the sparsity s of θ_τ and β_τ is below this truncation parameter the estimation will adapt to this more favorable design. □

2.2. **Unknown Conditional Density Function.** The implementation of the algorithms in Section 2.1 requires the knowledge of the conditional density function f_i which is typically unknown and needs to be estimated (under heteroscedasticity). Following [22] and letting $Q(\cdot \mid d_i, z_i)$ denote the conditional quantile function of the outcome, we shall use the observation that

$$f_i = \frac{1}{\partial Q(\tau \mid d_i, z_i)/\partial\tau}$$

to estimate f_i. Letting $\widehat{Q}(u \mid z_i, d_i)$ denote an estimate of the conditional u-quantile function $Q(u \mid z_i, d_i)$, based on ℓ_1-penalized quantile regression or the associated post-selection estimator, and $h = h_n \to 0$ denote a a bandwidth parameter, we let

$$\widehat{f}_i = \frac{2h}{\widehat{Q}(\tau + h \mid z_i, d_i) - \widehat{Q}(\tau - h \mid z_i, d_i)} \tag{2.14}$$

be an estimator of f_i. If the conditional quantile function is three times continuously differentiable, this estimator is based on the first order partial difference of the estimated conditional quantile function, and so it has the bias of order h^2.

It is also possible to use the following estimator:

$$\widehat{f}_i = h \left\{ \frac{3}{4}\{\widehat{Q}(\tau + h \mid z_i, d_i) - \widehat{Q}(\tau - h \mid z_i, d_i)\} - \frac{1}{12}\{\widehat{Q}(\tau + 2h \mid z_i, d_i) - \widehat{Q}(\tau - 2h \mid z_i, d_i)\} \right\}^{-1}, \tag{2.15}$$

which has the bias of order h^4 under additional smoothness assumptions. We denote by $\mathcal{U}$ the finite set of quantile indices used in the estimation of the conditional density.

Under mild regularity conditions the estimators (2.14) and (2.15) achieve

$$\widehat{f}_i - f_i = O\left(h^{\bar{k}} + \max_{u \in \mathcal{U}} \frac{|\widehat{Q}(\tau + u \mid d_i, z_i) - \widehat{Q}(\tau - u \mid d_i, z_i)|}{h} \right), \tag{2.16}$$

where $\bar{k} = 2$ for (2.14) and $\bar{k} = 4$ for (2.15).

Then Algorithms 1 and 2 are modified by replacing f_i with $\widehat{f}_i$.

Algorithm 1′ (Instrumental Quantile Regression with Optimal Instrument)

(1) Run ℓ_1-penalized quantile regressions of y_i on d_i and x_i to compute $(\widehat{\alpha}_u, \widehat{\beta}_u^{(2s)})$, $u \in \{\tau\} \cup \mathcal{U}$.
(2) Compute $\widehat{f}_i$ and run Post-Lasso of $\widehat{f}_i d_i$ on $\widehat{f}_i x_i$ to compute the residual $\widetilde{v}_i := \widehat{f}_i(d_i - x_i'\widetilde{\theta}_\tau)$.
(3) Run Instrumental Quantile Regression of $y_i - x_i'\widetilde{\beta}_\tau$ on d_i using $\widetilde{v}_i$ as the instrument for d_i to compute $\check{\alpha}_\tau$, and set $\check{\beta}_\tau = \widetilde{\beta}_\tau$.

Algorithm 2′ (Weighted Quantile Regression after Double Selection)

(1) Run ℓ_1-penalized quantile regressions of y_i on d_i and x_i to compute $(\widehat{\alpha}_u, \widehat{\beta}_u^{(2s)})$, $u = \{\tau\} \cup \mathcal{U}$.
(2) Compute $\widehat{f}_i$ and run Lasso of $\widehat{f}_i d_i$ on $\widehat{f}_i x_i$ to compute $\widehat{\theta}_\tau$.
(3) Run quantile regression of $\widehat{f}_i y_i$ on $\widehat{f}_i d_i$ and $\{\widehat{f}_i x_{ij}, j \in \text{support}(\widehat{\beta}_\tau^{(2s)}) \cup \text{support}(\widehat{\theta}_\tau)\}$ to compute $(\check{\alpha}_\tau, \check{\beta}_\tau)$.

Comment 2.4 (Implementation of the estimates $\widehat{f}_i$)**.** There are several possible choices of tunning parameters to construct the estimates $\widehat{f}_i$, however, they need to be coordinated with the penalty level λ. Together with the recommendations made in Remark 2.3, we suggest to construct $\widehat{f}_i$ as in (2.14) with bandwidth $h := \min\{n^{-1/6}, \tau(1-\tau)/2\}$. Remark 3.4 below discusses in more detail the requirements associated with different choices for penalty level λ and bandwidth h. □

2.3. **Overview of Main Results on Estimation and Inference.** Under mild moment conditions and approximately sparsity assumptions, we established that the estimator $\check{\alpha}_\tau$, as defined in Algorithm 1′ or Algorithm 2′, is root-n consistent and asymptotically normal,

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) \rightsquigarrow N(0,1), \tag{2.17}$$

where $\sigma_n^2 = \tau(1-\tau)\bar{\mathrm{E}}[v_i^2]^{-1}$ is the semi-parametric efficiency bound for the partially linear quantile regression model. The convergence result holds under array asymptotics, permitting the data-generating process $\mathrm{P} = \mathrm{P}_n$ to change with n, which implies that these convergence results hold uniformly over substantive sets of data-generating processes. In particular, our approach and results do not require separation of regression coefficients away from zero (the so-called "beta-min" conditions) for their validity.

As a consequence, the confidence region defined as

$$\mathcal{C}_{\xi,n} := \{\alpha \in \mathbb{R} : |\alpha - \check{\alpha}_\tau| \leqslant \widehat{\sigma}_n \Phi^{-1}(1-\xi/2)/\sqrt{n}\} \tag{2.18}$$

has asymptotic coverage of $1-\xi$ provided the estimate $\widehat{\sigma}_n^2$ is consistent for σ_n^2, namely $\widehat{\sigma}_n^2/\sigma_n^2 = 1 + o_{\mathrm{P}}(1)$. These confidence regions are asymptotically valid uniformly over a large class of data-generating processes P_n.

There are different possible choices of estimators for σ_n:

$$\begin{gathered}\widehat{\sigma}_{1n}^2 := \tau(1-\tau)\mathbb{E}_n[\widetilde{v}_i^2]^{-1}, \quad \widehat{\sigma}_{2n}^2 := \tau(1-\tau)\{\mathbb{E}_n[\widehat{f}_i^2(d_i, x_{i\check{T}}')'(d_i, x_{i\check{T}}')]\}_{11}^{-1}, \\ \widehat{\sigma}_{3n}^2 := \mathbb{E}_n[\widehat{f}_i d_i \widetilde{v}_i]^{-2}\mathbb{E}_n[(1\{y_i \leqslant d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\} - \tau)^2\widetilde{v}_i^2],\end{gathered} \tag{2.19}$$

where $\check{T} = \mathrm{support}(\widehat{\beta}_\tau) \cup \mathrm{support}(\widehat{\theta}_\tau)$ is the set of controls used in the double selection quantile regression. Although all three estimates are consistent under similar regularities conditions, their finite sample behaviour might differ. Based on the small-sample performance in computational experiments, we recommend the use of $\widehat{\sigma}_{3n}$ for the optimal IV estimator and $\widehat{\sigma}_{2n}$ for the double selection estimator.

Additionally, the criterion function of the instrumental quantile regression is the Neyman-type score statistic

$$L_n(\alpha) = \frac{|\mathbb{E}_n[\varphi_\tau(y_i, x_i'\check{\beta}_\tau + d_i\alpha)\widetilde{v}_i]|^2}{\mathbb{E}_n[\{\varphi_\tau(y_i, x_i'\check{\beta}_\tau + d_i\alpha)\widetilde{v}_i\}^2]},$$

is asymptotically distributed as chi-squared with 1 degree of freedom, when evaluated at the true value $\alpha = \alpha_\tau$, namely

$$nL_n(\alpha_\tau) \rightsquigarrow \chi^2(1). \tag{2.20}$$

The convergence result also holds under array asymptotics, permitting the data-generating process $\mathrm{P} = \mathrm{P}_n$ to change with n, which implies that these convergence results hold uniformly over substantive sets of data-generating processes. In particular, this result does not rely on the so-called beta-min conditions for its validity. This property allows the construction of another confidence region:

$$\mathcal{I}_{\xi,n} := \{\alpha \in \mathcal{A}_\tau : nL_n(\alpha) \leqslant (1-\xi) - \text{quantile of } \chi^2(1)\}, \tag{2.21}$$

which has asymptotic coverage level of $1-\xi$. These confidence regions too are asymptotically valid uniformly over a large class $\mathcal{P}_n$ of data-generating processes P_n.

3. Main results

In this section we provide sufficient conditions and formally state the main results of the paper.

3.1. **Regularity Conditions.**

Here we provide regularity conditions that are sufficient for validity of the main estimation and inference result. Throughout the paper, we let c, C, and q be absolute constants, and let $\ell_n \nearrow \infty, \delta_n \searrow 0$, and $\Delta_n \searrow 0$ be sequences of absolute positive constants.

We assume that for each n the following condition holds on the data generating process $\mathrm{P} = \mathrm{P}_n$.

Condition AS (P). *(i) Let $(z_i)_{i=1}^n$ denote a non-stochastic sequence and P denote a dictionary of transformations of z_i, which may depend on n but not on* P*. The p-dimensional vector $x_i = P(z_i)$ of covariates are normalized so that $\mathbb{E}_n[x_{ij}^2] = 1$, $j = 1, \ldots, p$, and $\{(y_i, d_i) : i = 1, \ldots, n\}$ be independent random vectors that obey the model given by (2.1) and (2.7) (ii) Functions g_τ and m_τ admit an approximately sparse form. Namely there exists $s \geqslant 1$ and β_τ and θ_τ, which depend on n and* P*, such that*

$$m_\tau(z_i) = x_i'\theta_\tau + r_{\theta\tau i}, \quad \|\theta_\tau\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{\theta\tau i}^2]\}^{1/2} \leqslant C\sqrt{s/n}, \tag{3.22}$$

$$g_\tau(z_i) = x_i'\beta_\tau + r_{g\tau i}, \quad \|\beta_\tau\|_0 \leqslant s, \quad \{\mathbb{E}_n[r_{g\tau i}^2]\}^{1/2} \leqslant C\sqrt{s/n}. \tag{3.23}$$

(iii) The conditional distribution function of ϵ_i is absolutely continuous with continuously differentiable density $f_{\epsilon_i}(\cdot \mid d_i, z_i)$ such that $0 < \underline{f} \leqslant f_i \leqslant \sup_\epsilon f_{\epsilon_i|d_i,z_i}(\epsilon \mid d_i, z_i) \leqslant \bar{f}$, $\sup_\epsilon |f'_{\epsilon_i|d_i,z_i}(\epsilon \mid d_i, z_i)| < \bar{f}'$. (iv) The following moment conditions apply: $|\bar{\mathrm{E}}[f_i v_i r_{g\tau i}]| \leqslant \delta_n n^{-1/2}$, $\bar{\mathrm{E}}[d_i^8] + \bar{\mathrm{E}}[v_i^8] \leqslant C$, $c \leqslant \mathrm{E}[v_i^2 \mid z_i] \leqslant C$ a.s. $1 \leqslant i \leqslant n$, $\max_{1\leqslant j\leqslant p}\{\bar{\mathrm{E}}[x_{ij}^2 d_i^2] + \bar{\mathrm{E}}[|x_{ij}^3 v_i^3|]\} \leqslant C$. (v) We have that $K_x = \max_{i\leqslant n}\|x_i\|_\infty$, $K_x^q \log p \leqslant \delta_n n$ for some $q > 4$, and s satisfies $(K_x^2 s^2 + s^3)\log^3(p \vee n) \leqslant n\delta_n$.

Condition AS(i) imposes the setting discussed in Section 2 in which the ϵ_i error term has zero τ-conditional quantile. The approximate sparsity condition AS(ii) is the main assumption for establishing the key inferential result. Condition AS(iii) is a standard assumption on the conditional density function in the quantile regression literature see [22] and the instrumental quantile regression literature [15]. Condition AS(iv) imposes some moment conditions. Condition AS(v) imposes growth conditions on s, p, K_x and n.

The next condition concerns the behavior of the Gram matrix $\mathbb{E}_n[x_i x_i']$. Whenever $p > n$, the empirical Gram matrix $\mathbb{E}_n[x_i x_i']$ does not have full rank and in principle is not well-behaved. However, we only need good behavior of smaller submatrices. Define the minimal and maximal m-sparse eigenvalue of a semi-definite matrix M as

$$\phi_{\min}(m)[M] := \min_{1\leqslant\|\delta\|_0\leqslant m} \frac{\delta' M \delta}{\|\delta\|^2} \quad \text{and} \quad \phi_{\max}(m)[M] := \max_{1\leqslant\|\delta\|_0\leqslant m} \frac{\delta' M \delta}{\|\delta\|^2}. \tag{3.24}$$

To assume that $\phi_{\min}(m)[M] > 0$ requires that all m by m submatrices of M are positive definite. We shall employ the following condition as a sufficient condition for our results.

Condition SE (P). *The maximal and minimal $\ell_n s$-sparse eigenvalues are bounded from below and away from zero, namely with probability at least $1-\Delta_n$, for $\tilde{x}_i = [d_i, x_i']'$*

$$\kappa' \leqslant \phi_{\min}(\ell_n s)[\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']] \leqslant \phi_{\max}(\ell_n s)[\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']] \leqslant \kappa'',$$

where $0 < \kappa' < \kappa'' < \infty$ are absolute constants.

For notational convenience we write $\phi_{\min}(m) := \phi_{\min}(m)[\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']]$ and $\phi_{\max}(m) := \phi_{\max}(m)[\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']]$. It is well-known that the first part of Condition SE is quite plausible for many designs of interest. For instance, Theorem 3.2 in [32] (see also [41] and [1]) shows that Condition SE holds for i.i.d. zero-mean sub-Gaussian regressors and $s(\log n)(\log p)/n \leqslant \delta_n \to 0$; while Theorem 1.8 [32] (see also Lemma 1 in [4]) shows that Condition SE holds for i.i.d. bounded zero-mean regressors with $\|x_i\|_\infty \leqslant K_x$ a.s. $K_x^2 s(\log^3 n)\{\log(p \vee n)\}/n \leqslant \delta_n \to 0$.

3.2. **Main results for the case with known density.** In this section we begin to state our theoretical results for the case where density values f_i are either known or constant and unknown. The case of constant density $f_i = f$ arises under conditional homoscedasticity, and in this case any constant value can be used as an "estimate", since it cancels in the definition of the estimators in Algorithms 1 and 2. Hence the results of this section are practically useful in homoscedastic cases; otherwise, they serve as a theoretical preparation of the results for the next subsection, where the unknown densities f_i will be estimated.

We first show that the optimal IV estimator based on Algorithm 1 with parameters (2.11)-(2.13) is root-n consistent and asymptotically normal.

Theorem 1 (Optimal IV estimator, conditional density f_i is known)**.** *Let $\{\mathrm{P}_n\}$ be a sequence of data-generating processes. Assume conditions AS (P) and SE (P) hold for $\mathrm{P} = \mathrm{P}_n$ for each n. Then, the optimal IV estimator $\check{\alpha}_\tau$ and the L_n function based on Algorithm 1 with parameters (2.11)-(2.13) obeys as $n \to \infty$*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) = \mathbb{U}_n(\tau) + o_P(1) \quad \mathbb{U}_n(\tau) \rightsquigarrow N(0,1)$$

where $\sigma_n^2 = \tau(1-\tau)\bar{\mathrm{E}}[v_i^2]^{-1}$ and

$$\mathbb{U}_n(\tau) := \frac{\{\tau(1-\tau)\bar{\mathrm{E}}[v_i^2]\}^{-1/2}}{\sqrt{n}} \sum_{i=1}^n (\tau - 1\{U_i \leqslant \tau\})v_i$$

where $U_1, \ldots, U_n$ are i.i.d. uniform $(0,1)$ random variables, independently distributed of $v_1, \ldots, v_n$. Furthermore,

$$nL_n(\alpha_\tau) = \mathbb{U}_n^2(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n^2(\tau) \rightsquigarrow \chi^2(1).$$

Theorem 1 relies on post model selection estimators which in turn relies on achieving sparse estimates $\widehat{\beta}_\tau$ and $\widehat{\theta}_\tau$. The sparsity of $\widehat{\theta}_\tau$ is derived in Section A.2 under the recommended penalty choices. The sparsity of $\widehat{\beta}_\tau$ is not guaranteed under the recommended choices of penalty level λ_τ which leads to sharp rates. We ensure sparsity by truncating to zero the smallest components. Lemma 6 shows that such operation does not impact the rates of convergence provided the largest $2s$ non-zero components are preserved.

We also establish a similar result for the double selection estimator based on Algorithm 2 with parameters (2.11)-(2.12).

Theorem 2 (Weighted double selection, known conditional density f_i)**.** *Let $\{\mathrm{P}_n\}$ be a sequence of data-generating processes. Assume conditions AS(P) and SE(P) hold for $\mathrm{P} = \mathrm{P}_n$ for each n. Then, the double selection estimator $\check{\alpha}_\tau$ and the L_n function based on Algorithm 2 with parameters (2.11)-(2.12) obeys as $n \to \infty$*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) = \mathbb{U}_n(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n(\tau) \rightsquigarrow N(0,1)$$

where $\sigma_n^2 = \tau(1-\tau)\bar{\mathrm{E}}[v_i^2]^{-1}$ and

$$\mathbb{U}_n(\tau) := \frac{\{\tau(1-\tau)\bar{\mathrm{E}}[v_i^2]\}^{-1/2}}{\sqrt{n}} \sum_{i=1}^{n} (\tau - 1\{U_i \leqslant \tau\})v_i$$

where $U_1, \ldots, U_n$ are i.i.d. uniform $(0,1)$ random variables, independently distributed of $v_1, \ldots, v_n$. Furthermore,

$$nL_n(\alpha_\tau) = \mathbb{U}_n^2(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n^2(\tau) \rightsquigarrow \chi^2(1).$$

Importantly, the results in Theorems 1 and 2 allows for the data generating process to depend on the sample size n and have no requirements on the separation from zero of the coefficients. In particular these results allow for sequences of data generating processes for which perfect model selection is not possible. In turn this translates into uniformity properties over a large class of data generating process. Next we formalize these uniform properties. We let $\mathcal{P}_n$ the collection of distributions P for the data $\{(y_i, d_i, z_i')'\}_{i=1}^n$ such that Conditions AS(P) and SE(P) hold for the given n. This is the collection of all approximately sparse models where the stated above sparsity conditions, moment conditions, and growth conditions hold.

Corollary 1 (**Uniform $\sqrt{n}$-Rate of Consistency and Uniform Normality**)**.** *Let $\mathcal{P}_n$ be the collection of all distributions of $\{(y_i, d_i, z_i')'\}_{i=1}^n$ for which Conditions AS and SE are satisfied for the given $n \geqslant 1$. Then either the optimal IV or the double selection estimator, $\check{\alpha}_\tau$, are $\sqrt{n}$-consistent and asymptotically normal uniformly over $\mathcal{P}_n$, namely*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{P}_n} \sup_{t\in\mathbb{R}} |\mathrm{P}(\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) \leqslant t) - \mathrm{P}(N(0,1) \leqslant t)| = 0.$$

Corollary 2 (**Uniform Validity of Confidence Regions**)**.** *Let $\mathcal{P}_n$ be the collection of all distributions of $\{(y_i, d_i, z_i')'\}_{i=1}^n$ for which Conditions AS and SE are satisfied for the given $n \geqslant 1$. Then the confidence regions $\mathcal{C}_{\xi,n}$ and $\mathcal{I}_{\xi,n}$ defined based on either the optimal IV estimator or by the double selection estimator are asymptotically valid uniformly in n, that is*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{P}_n} |\mathrm{P}(\alpha_\tau \in \mathcal{C}_{\xi,n}) - (1-\xi)| = 0 \quad \textit{and} \quad \lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{P}_n} |\mathrm{P}(\alpha_\tau \in \mathcal{I}_{\xi,n}) - (1-\xi)| = 0.$$

The uniformity results for the approximately sparse and heteroscedastic case are new even under fixed p asymptotics.

Comment 3.1. Both algorithms assume that the values of the conditional density function f_i, $i = 1, \ldots, n$, are known. In fact it suffices to know them up to a multiplicative constant, which allows to

cover the homoscedastic case, where $f_i = 1, i = 1, \ldots, n$. In heteroscedastic settings we shall need to estimate f_i, and we analyze this case in the next subsection. □

Comment 3.2 (Inference Based on the Pivotal Process $\mathbb{U}_n(\tau)$)**.** In addition to the asymptotic normality, Theorems 1 and 2 establish that the rescaled estimation error $\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau)$ is approximately equal to the process $\mathbb{U}_n(\tau)$, which is pivotal conditional on $v_1, \ldots, v_n$. Such property is very useful since it is easy to simulate $\mathbb{U}_n(\tau)$ conditional on $v_1, \ldots, v_n$. Thus this representation provides us with another procedure to construct confidence intervals that does not rely on asymptotic normality.

3.3. **Main Results for the case of unknown density.** Next we provide formal results to the case the conditional probability density function is unknown. In this case it is necessary to estimate the weights f_i, and this estimation has a non-trivial impact on the analysis. Condition D summarizes sufficient conditions to account for the impact of the density estimation.

Condition D. *(i) For a bandwidth h, assume that $g_u(z_i) = x_i'\beta_u + r_{ui}$ where the approximation errors satisfy $\bar{\mathrm{E}}[r_{ui}^2] \leqslant \delta_n n^{-1/2}$ and $|r_{ui}| \leqslant \delta_n h$ for all $i = 1, \ldots, n$, and the vector β_u satisfies $\|\beta_u\|_0 \leqslant s$, for $u = \tau, \tau \pm h, \tau \pm 2h$. (ii) Suppose $\|\widehat{\beta}_u\|_0 \leqslant Cs$ and $\|d_i\widehat{\alpha}_u + x_i'\widehat{\beta}_u - g_{ui} - d_i\alpha_u\|_{2,n} \leqslant C\sqrt{s\log(p\vee n)/n}$ with probability at least $1 - \Delta_n$ for $u = \tau, \tau \pm h, \tau \pm 2h$. (iii) $K_x^2 s^2 \log(p\vee n) \leqslant \delta_n n h^2$, $h^{\bar{k}}\sqrt{s\log p} \leqslant \delta_n, h^{\bar{k}-1}\sqrt{s\log p}(\sqrt{n\log p}/\lambda) \leqslant \delta_n, h^{2\bar{k}}\sqrt{n}(\sqrt{n\log p}/\lambda) \leqslant \delta_n, s^2\log^2 p \leqslant \delta_n n h^2, s^2\log^3 p \leqslant \delta_n h^4\lambda^2, \lambda s\sqrt{\log p} \leqslant \delta_n n$ (iv) For $s_{\theta\tau} = s + \frac{ns\log(n\vee p)}{h^2\lambda^2} + \left(\frac{nh^{\bar{k}}}{\lambda}\right)^2$, we have $0 < \kappa' < \phi_{\min}(\ell_n s_{\theta\tau}) \leqslant \phi_{\max}(\ell_n s_{\theta\tau}) \leqslant \kappa'' < \infty$ with probability $1 - \Delta_n$.*

Comment 3.3. Condition D(i) imposes the approximately sparse assumption for the u-conditional quantile function for quantile indices u in a neighborhood of the quantile index τ. Condition D(ii) is a high level condition on the estimates of β_u which are typically satisfied by ℓ_1-penalized quantile regression estimators. As before sparsity can be achieved by truncating these vectors. Condition D(iii) provide growth conditions relating s, p, n, h and λ. Remark 3.4 below discusses specific choices of penalty level λ and of bandwidth h together with the implied conditions on the triple (s, p, n). □

Next we establish the main inferential results for the case with estimated conditional density weights. We begin with the optimal IV estimator which is based on Algorithm 1′ with parameters λ_τ as in (2.11), $\widehat{\Gamma}_\tau$ as in (2.12) with f_i replaced with $\widehat{f}_i$, and $\mathcal{A}_\tau$ as in (2.13). The choices of λ and h satisfy Condition D.

Theorem 3 (Optimal IV estimator, estimated conditional density $\widehat{f}_i$)**.** *Let $\{\mathrm{P}_n\}$ be a sequence of data-generating processes. Assume conditions AS (P) and D (P) hold for $\mathrm{P} = \mathrm{P}_n$ for each n. Then, the optimal IV estimator $\check{\alpha}_\tau$ and the L_n function based on Algorithm 3 with parameters (2.11)-(2.13) obeys as $n \to \infty$*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) = \mathbb{U}_n(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n(\tau) \rightsquigarrow N(0,1)$$

where $\sigma_n^2 = \tau(1-\tau)\bar{\mathrm{E}}[v_i^2]^{-1}$ and

$$\mathbb{U}_n(\tau) := \frac{\{\tau(1-\tau)\bar{\mathrm{E}}[v_i^2]\}^{-1/2}}{\sqrt{n}}\sum_{i=1}^n(\tau - 1\{U_i \leqslant \tau\})v_i$$

where $U_1, \ldots, U_n$ are i.i.d. uniform $(0,1)$ random variables, independently distributed of $v_1, \ldots, v_n$. Furthermore,

$$nL_n(\alpha_\tau) = \mathbb{U}_n^2(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n^2(\tau) \rightsquigarrow \chi^2(1).$$

The result continues to apply if σ_n^2 is replaced by any of the estimators in (2.19), namely $\widehat{\sigma}_{kn}/\sigma_n = 1 + o_{\mathrm{P}}(1)$ for $k = 1, 2, 3$.

The following is a corresponding result for the double selection estimator based on Algorithm 2′ with parameters λ_τ as in (2.11), and $\widehat{\Gamma}_\tau$ as in (2.12) with f_i replaced with $\widehat{f}_i$. As before the choices of λ and h satisfy Condition D and are discussed in detail below.

Theorem 4 (Weighted double selection estimator, estimated conditional density $\widehat{f}_i$)**.** *Let $\{\mathrm{P}_n\}$ be a sequence of data-generating processes. Assume conditions AS(P) and D(P) hold for $\mathrm{P} = \mathrm{P}_n$ for each n. Then, the double selection estimator $\check{\alpha}_\tau$ and the L_n function based on Algorithm 4 with parameters (2.11)-(2.12) obeys as $n \to \infty$*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) = \mathbb{U}_n(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n(\tau) \rightsquigarrow N(0,1)$$

where $\sigma_n^2 = \tau(1-\tau)\bar{\mathrm{E}}[v_i^2]^{-1}$ and

$$\mathbb{U}_n(\tau) := \frac{\{\tau(1-\tau)\bar{\mathrm{E}}[v_i^2]\}^{-1/2}}{\sqrt{n}} \sum_{i=1}^{n} (\tau - 1\{U_i \leqslant \tau\})v_i$$

where $U_1, \ldots, U_n$ are i.i.d. uniform $(0,1)$ random variables, independently distributed of $v_1, \ldots, v_n$. Furthermore,

$$nL_n(\alpha_\tau) = \mathbb{U}_n^2(\tau) + o_P(1) \quad \textit{and} \quad \mathbb{U}_n^2(\tau) \rightsquigarrow \chi^2(1).$$

The result continues to apply if σ_n^2 is replaced by any of the estimators in (2.19), namely $\widehat{\sigma}_{kn}/\sigma_n = 1 + o_{\mathrm{P}}(1)$ for $k = 1, 2, 3$.

Comment 3.4 (Choice of Bandwidth h and Penalty Level λ in Step 2)**.** The proofs of Theorems 3 and 4 provide a detailed analysis for generic choice of bandwidth h and the penalty level λ in Step 2 under Condition D. Here we discuss two particular choices: for $\gamma = 0.05/\{n \vee p \log n\}$

$$\begin{array}{ll} \text{(i)} & \lambda = h^{-1}\sqrt{n}\Phi^{-1}(1-\gamma), \\ \text{(ii)} & \lambda = 1.1\sqrt{n}2\Phi^{-1}(1-\gamma). \end{array}$$

The choice (i) for λ leads to the optimal prediction rate by adjusting to the slower rate of convergence of $\widehat{f}_i$, see (2.16). The choice (ii) for λ corresponds to the (standard) choice of penalty level in the literature for Lasso. For these choices Condition D(iii) simplifies to

$$\begin{array}{ll} \text{(i)} & h^{\bar{k}}\sqrt{s\log p} \leqslant \delta_n, \quad h^{2\bar{k}+1}\sqrt{n} \leqslant \delta_n, \quad \text{and} \quad K_x^2 s^2 \log^2(p \vee n) \leqslant \delta_n n h^2, \\ \text{(ii)} & h^{\bar{k}-1}\sqrt{s\log p} \leqslant \delta_n, \quad h^{2\bar{k}}\sqrt{n} \leqslant \delta_n, \quad \text{and} \quad s^2 \log^2 p \leqslant \delta_n n h^4. \end{array}$$

For example, using the choice of $\widehat{f}_i$ as in (2.15) so that $\bar{k} = 4$, we have that the following choice growth conditions suffice for the conditions above:

$$\begin{array}{ll} \text{(i)} & K_x^3 s^3 \log^3(p \vee n) \leqslant \delta_n n \text{ and } h = n^{-1/6} \\ \text{(ii)} & (s\log(p \vee n) + K_x^3) s^3 \log^3(p \vee n) \leqslant \delta_n n, \text{ and } h = n^{-1/8} \end{array}$$

□

4. Empirical Performance

We present monte-carlo experiments, followed by a data-analytic example.

4.1. **Monte-Carlo Experiments.** In this section we provide a simulation study to assess the finite sample performance of the proposed estimators and confidence regions. We shall focus on examining the inferential properties of the confidence regions based upon Algorithms $1'$ and $2'$, and contrast them with the confidence intervals based on naive (standard) selection.

We considered the following regression model for $\tau = 1/2$:

$$y = d\alpha_\tau + x'(c_y\nu_0) + \epsilon, \quad \epsilon \sim N(0, \{2 - \mu + \mu d^2\}/2), \tag{4.25}$$

$$d = x'(c_d\nu_0) + \tilde{v}, \quad \tilde{v} \sim N(0, 1), \tag{4.26}$$

where $\alpha_\tau = 1/2$, $\theta_{0j} = 1/j^2, j = 1, \ldots, p$, $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Sigma)$, and the errors ϵ and $\tilde{v}$ are independent. In this case, the optimal instrument is $v = \tilde{v}/\{\sqrt{\pi(2 - \mu + \mu d^2)}\}$. 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 coefficient $\mu \in \{0, 1\}$ which makes the conditional density function of ϵ homoscedastic if $\mu = 0$ and heteroscedastic if $\mu = 1$. The coefficients c_y and c_d are used to control the R^2 in the equations: $y - d\alpha_\tau = x'(c_y\nu_0) + \epsilon$ and $d = x'(c_d\nu_0) + \tilde{v}$; we denote the values of R^2 in each equation by R_y^2 and R_d^2. We consider values (R_y^2, R_d^2) in the set $\{0, .1, .2, \ldots, .9\} \times \{0, .1, .2, \ldots, .9\}$. Therefore we have 100 different designs and we perform 500 Monte-Carlo repetitions for each design. For each repetition we draw new vectors x_i's and errors ϵ_i's and $\tilde{v}_i$'s.

The design above with $g_\tau(z) = x'(c_y\nu_0)$ is an approximately sparse model; and the gradual decay of the components of ν_0 rules out typical "separation from zero" assumptions of the coefficients of "important" covariates. Thus, we anticipate that inference procedures which rely on the model selection in the direct equation (4.25) only will not perform well in our simulation study. We refer to such selection procedures as the "naive"/single selection and the call the resulting inference procedures the post "naive"/single selection inference. To be specific, in our simulation study, the "naive" selection procedure applies ℓ_1-penalized τ-quantile regression of y on d and x to select a subset of covariates that have predictive power for y, and then runs $\tau-$quantile regression of y on d and the selected covariates, omitting the covariates that were not selected. This procedure is the standard procedure that is often employed in practice.

The model in (4.25) can be heteroscedastic, since when $\mu \neq 0$ the distribution of the error term might depend on the main regressor of interest d. Under heteroscedasticity, our procedures require estimations of the conditional probability density function f_i, and we do so via (2.14). We perform estimation of f_i's even in the homoscedastic case ($\mu = 0$), since we do not want rely on whether the assumption of homoscedasticity is valid or not. In other words, we use Algorithms $1'$ and $2'$ in both heteroscedastic and homoscedastic cases. We use $\widehat{\sigma}_{3n}$ as the standard error for the optimal IV estimator, and $\widehat{\sigma}_{2n}$ as the standard error for the post double selection estimator. As a benchmark we consider the standard

post-model selection procedure based on ℓ_1-penalized quantile regression method (post single selection) based upon equation (4.25) alone, as define in the previous paragraph.

In Figure 1 we report the results for the homoscedastic case ($\mu = 0$). In our study, we focus on the quality of inferential procedures – namely on the rejection frequency of the confidence intervals with the nominal coverage probability of 95%, and so the figure reports these frequencies. Ideally we should see the rejection rate of 5%, the nominal level, regardless of what the underlying generating process $\mathrm{P} \in \mathcal{P}_n$ is. The is the so called uniformity property or honesty property of the confidence regions (see, e.g., Romano and Wolf [31], Romano and Shaikh [30], and Leeb and Pötscher [27]). The top left plot of Figure 1 reports the empirical rejection probabilities for the naive post single selection procedure. These empirical rejection probabilities deviate strongly away from the nominal level of 5%, demonstrating the striking lack of robustness of this standard method. This is perhaps expected due to the Monte-Carlo design having regression coefficients not well separated from zero (that is, "beta min" condition does not hold here). In sharp contrast, we see from top right and bottom right and left plots of Figure 1, that both of our proposed procedures perform substantially better, yielding empirical rejection probabilities close to the desired nominal level of 5%. We also see from comparing the bottom left plot to other plots that the confidence regions based on the post-double selection method somewhat outperform the optimal IV estimator.

Figure 2 we report the results for the heteroscedastic case ($\mu = 1$). The figure displays the (empirical) rejection probability of the confidence intervals with nominal coverage of 95%. As before, ideally we should see the empirical rejection probability of 5%. Again the top left figure reports the results for the confidence intervals based on the naive post model selection estimator. Here too we see the striking lack of robustness of this standard method; this occurs due to the direct equation (4.25) having coefficients ν_0 that are not well separated from zero. We see from top right and bottom right and left plots of Figure 1, that both of our proposed procedures perform substantially better, however, the optimal IV procedure does not do as well as in the homoscedastic case. We also see from comparing the bottom left plot to other plots that the confidence regions based on the post-double selection method significantly outperform the optimal IV estimator, yielding empirical rejection frequencies close to the nominal level of 5%.

Thus, based on these experiments, we recommend to use the post-double selection procedure over the optimal IV procedure.

4.2. **Inference on Risk Factors in Childhood Malnutrition.** The purpose of this section is to examine practical usefulness of the new methods and contrast them with the standard post-selection inference (that assumes that selection had worked perfectly).

We will assess statistical significance of socio-economic and biological factors on children's malnutrition, providing a methodological follow up on the previous studies done by [17] and [21]. The measure of malnutrition is represented by the child's height, which will be our response variable y. The socio-economic and biological factors will be our regressors x, which we shall describe in more detail below. We shall estimate the conditional first decile function of the child's height given the factors (that is, we set $\tau = .1$). We'd like to perform inference on the size of the impact of the various factors on the conditional

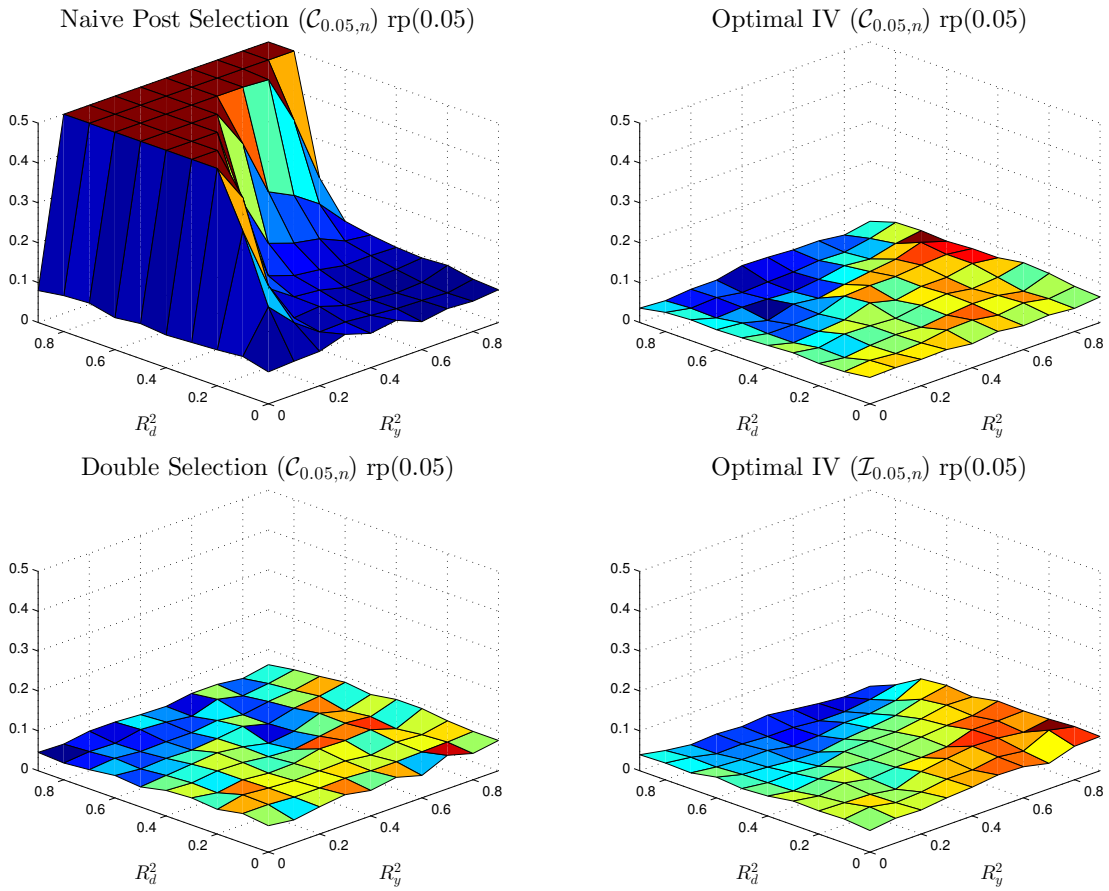

FIGURE 1. For the homoscedastic design ($\mu = 0$), the figure displays the rejection probabilities of the following confidence regions with nominal coverage of 95%: (a) the confidence region based upon naive (single) selection procedure (top left panel), (b) the confidence region $\mathcal{C}_{0.05,n}$ based the optimal IV estimator based on (top right), (c) the confidence region, as defined in Algorithm 1′, $\mathcal{I}_{0.05,n}$ based on the optimal IV procedure (bottom right panel), as defined in Algorithm 1′, and (d) the confidence region $\mathcal{C}_{0.05,n}$ based on the post double selection estimator (bottom left panel), as defined in Algorithm 1′. Each point in each of the plots corresponds to a different data-generating process indexed by pairs of R^2 values (R_d^2, R_y^2) varying over the set $\{0, .1, \ldots, .9\} \times \{0, .1, \ldots, .9\}$. The results are based on 500 replications for each of the 100 combinations of R^2's in each equation. The ideal rejection probability should be 5%, so ideally we should be seeing a flat surface with height 5%.

decile of the child's height. The problem has material significance, so it is important to conduct statistical inference for this problem responsibly.

The data comes originally from the Demographic and Health Surveys (DHS) conducted regularly in more than 75 countries; we employ the same selected sample of 37,649 as in Koenker (2012). All children in the sample are between the ages of 0 and 5. The response variable y is the child's height in centimeters. The regressors x include child's age, breast feeding in months, mothers body-mass index (BMI), mother's age, mother's education, father's education, number of living children in the family, and a large number of categorical variables, with each category coded as binary (zero or one): child's gender (male or female),

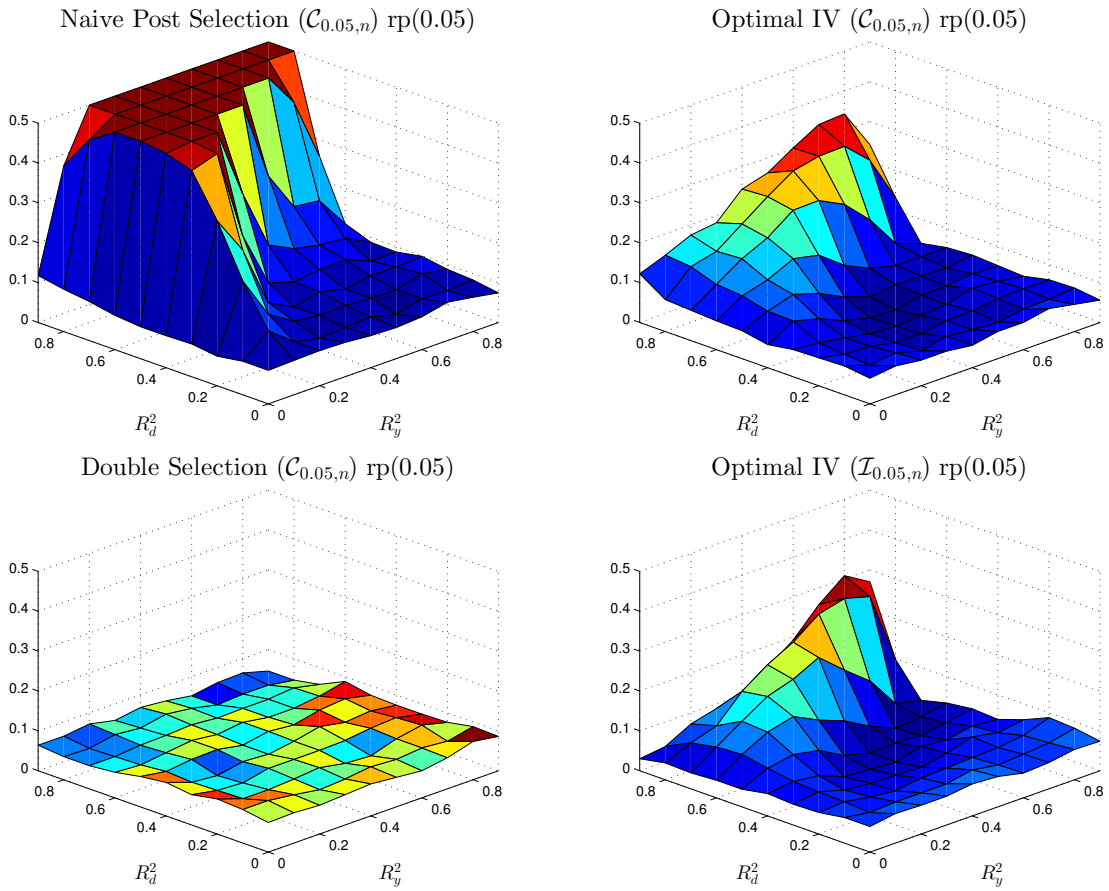

FIGURE 2. For the heteroscedastic design ($\mu = 1$), the figure displays the rejection probabilities of the following confidence regions with nominal coverage of 95%: (a) the confidence region based upon naive (single) selection procedure (top left panel), (b) the confidence region $\mathcal{C}_{0.05,n}$ based the optimal IV estimator based on (top right), (c) the confidence region, as defined in Algorithm 1′, $\mathcal{I}_{0.05,n}$ based on the optimal IV procedure (bottom right panel), as defined in Algorithm 1′, and (d) the confidence region $\mathcal{C}_{0.05,n}$ based on the post double selection estimator (bottom left panel), as defined in Algorithm 1′. Each point in each of the plots corresponds to a different data-generating process indexed by pairs of R^2 values (R_d^2, R_y^2) varying over the set $\{0, .1, \ldots, .9\} \times \{0, .1, \ldots, .9\}$. The results are based on 500 replications for each of the 100 combinations of R^2's in each equation. The ideal rejection probability should be 5%, so ideally we should be seeing a flat surface with height 5%.

twin status (single or twin), the birth order (first, second, third, fourth, or fifth), the mother's employment status (employed or unemployed), mother's religion (Hindu, Muslim, Christian, Sikh, or other), mother's residence (urban or rural), family's wealth (poorest, poorer, middle, richer, richest), electricity (yes or no), radio (yes or no), television (yes or no), bicycle (yes or no), motorcycle (yes or no), and car (yes or no).

Although the number of covariates – 30 – is substantial, the sample size – 37,649– is much larger than the number of covariates. Therefore, the dataset is very interesting from a methodological point of

view, since it gives us an opportunity to compare various methods for performing inference to an "ideal" benchmark:

(1) The "ideal" benchmark here is the standard inference based on the standard quantile regression estimator without any model selection. Since the number of regressors p is much smaller than the sample size n, this is a very good option. The latter was proven theoretically in [18] and in [5] under the $p \to \infty, p^3/n \to 0$ regime. This is also the general option recommended by [22] and [26] in the fixed p regime. Note that this "ideal" option does not apply in practice when p is relatively large; however it certainly applies in the present example.

(2) The standard post-selection inference method is the existing benchmark. This method performs standard inference on the post-model selection estimator, "assuming" that the model selection had worked perfectly. While this approach has some justification, we expect it to perform poorly, based on our computational results and from theoretical results of [26]. In particular, it would be very interesting to see if it gives misleading results as compared to the "ideal" option.

(3) We propose two methods, one based on the instrumental regression estimator (Algorithm 1) and another based on double selection (Algorithm 2). The proposed methods do not assume perfect selection, but rather builds a protection against (moderate) model selection mistakes. From the theory we would expect the method to give results similar to the "ideal" option in (1).

We now will compare our proposal to the "ideal" benchmark and to the standard post-selection method. We report the empirical results in Table 4.2. The first column reports results for the option 1, reporting the estimates and standard errors enclosed in brackets. The second column reports results for option 2, specifically the point estimates resulting from the use of ℓ_1-penalized quantile regression and the post-penalized quantile regression, reporting the standard errors as if there had been no model selection. The third column and fourth column report the results for two versions – Algorithm 1 and Algorithm 2 – of option 3. Each column reports point estimates, the standard errors, and the confidence region obtained by inverting the robust L_n-statistic. Note that the Algorithms 1 and 2 are applied sequentially to each of the variables. Similarly, in order to provide estimates and confidence intervals for all variables using the naive approach, if a covariate was not selected by the ℓ_1-penalized quantile regression, it was included in the post-model selection quantile regression for that variable.

What we see is very interesting. First of all, let us compare "ideal" option (column 1) and the naive post-selection (column 2). Lasso selection method removes 16 out of 30 variables, many of which are highly significant, as judged by the "ideal" option. (To judge significance we use normal approximations and critical value of 3, which allows us to maintain 5% significance level after testing up to 50 hypotheses). In particular, we see that the following highly significant variables were dropped by Lasso: mother's BMI, mother's age, twin status, birth orders one and two, and indicator of the other religion. The standard post-model selection inference then makes the assumption that these are true zeros, which lead us to misleading conclusions about these effects. The standard post-model selection inference then proceeds to judge the significance of other variables, in some cases deviating sharply and significantly from the

"ideal" benchmark. For example, there is a sharp disagreement on magnitudes of the impact of the birth order variables and the wealth variables (for "richer" and "richest" categories). Overall, for the naive post-selection, 8 out of 30 coefficients were more than 3 standard errors away from the coefficients of the "ideal" option.

We now proceed to compare our proposed options to the "ideal" option. We see approximate agreement in terms of magnitude, signs of coefficients, and in standard errors. In few instances, for example, for the car ownership regressor, the disagreements in magnitude may appear large, but they become insignificant once we account for the standard errors. In particular, the pointwise 95% confidence regions constructed by inverting the L_n statistics all contain the estimates from the "ideal" option. Moreover, there is very little disagreement between Algorithms 1 (optimal IV) and Algorithm 2 (double selection). The agreement here is good news from the point of view of our theory, since it confirms what we had expected from our previous analysis. In particular, for the proposed methods, no coefficient estimate was more than 1.5 standard errors away from the coefficient of the "ideal" option.

The main conclusion from our study is that the standard/naive post-selection inference can give misleading results, confirming our expectations and confirming predictions of [26]. Moreover, the proposed inference procedures are able to deliver inference of high quality, which is very much in agreement with the "ideal" benchmark.

5. Discussion

5.1. **Variants of the Proposed Algorithms.** There are several different ways to implement the sequence of steps underlying the two procedures outlined in Algorithms 1 and 2. The estimation of the control function g_τ can be done through other regularization methods like ℓ_1-penalized quantile regression instead of the post-ℓ_1-penalized quantile regression estimator. The estimation of the instrument v in Step 2 can be carried out with Dantzig selector, square-root Lasso or the associated post-model selection could be used instead of Lasso or Post-Lasso. The instrumental quantile regression can be substituted by a 1-Step estimator from the ℓ_1-penalized quantile regression estimator $\widehat{\alpha}_\tau$ of the form $\check{\alpha}_\tau = \widehat{\alpha}_\tau + (\mathbb{E}_n[\widehat{v}_i^2])^{-1}\mathbb{E}_n[\varphi_\tau(y_i, \widehat{\alpha}_\tau d_i + x_i'\widehat{\beta}_\tau)\widehat{v}_i]$.

Other variants can be constructed by using another valid instrument. An instrument $\iota_i = \iota(d_i, z_i)$ is valid if it satisfies $\bar{\mathrm{E}}[f_i\iota_i \mid z_i] = 0$ and $\bar{\mathrm{E}}[f_i d_i \iota_i] \neq 0$. For example, a valid choice of instrument is $\iota_i = (d_i - \mathrm{E}[d_i \mid z_i])/f_i$. Typically this choice of instruments does not lead to a semi-parametric efficient estimator as the choices proposed in Algorithms 1 and 2 do. Nonetheless, the estimation of $\mathrm{E}[d_i \mid z_i]$ and f_i can be carried out separably which can lead to weaker regularity conditions.

5.2. **Uniform Inference over $\tau \in \mathcal{T}$ and Many Coefficients.** In some applications the interest lies on building confidence intervals for many coefficients simultaneously. Moreover, some applications would also be interest on a range of quantile indices. The methods developed here can be extended to the case $d \in \mathbb{R}^K$ and $\tau \in \mathcal{T}$

$$\tau - \text{quantile}(y \mid z, d) = \sum_{j=1}^{K} d_j \alpha_{\tau j} + \tilde{g}_\tau(z)$$

TABLE 1. Empirical Results

	(1)		(2)	(3)		
	quantile	ℓ_1-penalized	Naive	Optimal IV		Double Selection
Variable	regression	quantile regression	post selection	$\check{\alpha}_\tau$	$\mathcal{I}_{0.05,n}$	$\check{\alpha}_\tau$
cage	0.6456	0.6360	0.6458	0.6458	[0.6400, 0.6514]	0.6449
	(0.0030)		(0.0027)	(0.0025)		(0.0032)
mbmi	0.0603	—	0.0663	0.0550	[0.0132, 0.0885]	0.0582
	(0.0159)		(0.0139)	(0.0316)		(0.0173)
breastfeeding	0.0691	0.0538	0.0689	0.0689	[0.0577, 0.0762]	0.0700
	(0.0036)		(0.0038)	(0.0036)		(0.0044)
mage	0.0684	—	0.0454	0.0705	[0.0416, 0.0947]	0.0685
	(0.0090)		(0.0147)	(0.0109)		(0.0126)
medu	0.1590	0.2036	0.1870	0.1594	[0.1246, 0.1870]	0.1566
	(0.0136)		(0.0145)	(0.0153)		(0.0154)
edupartner	0.0175	0.0147	0.0460	0.0388	[0.0053, 0.0641]	0.0348
	(0.0125)		(0.0148)	(0.0143)		(0.0143)
deadchildren	-0.0680	—	-0.2121	-0.0791	[-0.3522, 0.0394]	-0.1546
	(0.1124)		(0.0978)	(0.0653)		(0.1121)
csexfemale	-1.4625	-1.0786	-1.5084	-1.5146	[-1.7166, -1.3322]	-1.5299
	(0.0948)		(0.0897)	(0.0923)		(0.1019)
ctwintwin	-1.7259	—	-1.8683	-1.8683	[-3.3481, -0.4652]	-1.9248
	(0.3741)		(0.2295)	(0.1880)		(0.7375)
cbirthorder2	-0.7256	—	-0.2230	-0.7408	[-1.0375, -0.3951]	-0.6818
	(0.1073)		(0.0983)	(0.1567)		(0.1337)
cbirthorder3	-1.2367	—	-0.5751	-1.0737	[-1.4627, -0.7821]	-1.1326
	(0.1315)		(0.1423)	(0.1556)		(0.1719)
cbirthorder4	-1.7455	-0.1892	-0.7910	-1.7219	[-2.2968, -1.2723]	-1.5819
	(0.2244)		(0.1938)	(0.2796)		(0.2193)
cbirthorder5	-2.4014	-0.8459	-1.1747	-2.3700	[-3.2407, -1.9384]	-2.3041
	(0.1639)		(0.1686)	(0.2574)		(0.2564)
munemployedemployed	0.0409	—	0.0077	0.0342	[-0.2052, 0.2172]	0.0379
	(0.1025)		(0.1077)	(0.1055)		(0.1124)
mreligionhindu	-0.4351	—	-0.2423	-0.5129	[-0.9171, -0.1523]	-0.5680
	(0.2232)		(0.1080)	(0.2277)		(0.1771)
mreligionmuslim	-0.3736	—	0.0294	-0.6177	[-1.1523, -0.1457]	-0.5119
	(0.2417)		(0.1438)	(0.2629)		(0.2176)
mreligionother	-1.1448	—	-0.6977	-1.2437	[-2.1037, -0.4828]	-1.1539
	(0.3296)		(0.3219)	(0.3390)		(0.3577)
mreligionsikh	-0.5575	—	0.3692	-0.5437	[-1.5591, 0.4243]	-0.3408
	(0.2969)		(0.1897)	(0.3653)		(0.3889)
mresidencerural	0.1545	—	0.1085	0.1519	[-0.1295, 0.3875]	0.1678
	(0.0994)		(0.1363)	(0.1313)		(0.1311)
wealthpoorer	0.2732	-0.0183	-0.1946	0.1187	[-0.1784, 0.5061]	0.2648
	(0.1761)		(0.1231)	(0.1505)		(0.1877)
wealthmiddle	0.8699	—	0.9197	0.9113	[0.4698, 1.3149]	0.9173
	(0.1719)		(0.2236)	(0.1784)		(0.2158)
wealthricher	1.3254	0.3252	0.5754	1.2751	[0.7515, 1.5963]	1.4040
	(0.2244)		(0.1408)	(0.1964)		(0.2505)
wealthrichest	2.0238	1.1167	1.2967	1.9149	[1.3086, 2.3893]	2.1133
	(0.2596)		(0.2263)	(0.2427)		(0.3318)
electricityyes	0.3866	0.3504	0.7555	0.4263	[0.1131, 0.7850]	0.4582
	(0.1581)		(0.1398)	(0.1572)		(0.1577)
radioyes	-0.0385	—	0.1363	0.0599	[-0.2100, 0.2682]	0.0640
	(0.1218)		(0.1214)	(0.1294)		(0.1207)
televisionyes	-0.1633	0.0122	-0.0774	-0.1112	[-0.3629, 0.0950]	-0.0880
	(0.1191)		(0.1234)	(0.0971)		(0.1386)
refrigeratoryes	0.1544	0.0899	0.2451	0.1907	[-0.1642, 0.5086]	0.2001
	(0.1774)		(0.2081)	(0.1716)		(0.1891)
bicycleyes	0.1438	—	0.1314	0.1791	[-0.0036, 0.3506]	0.1438
	(0.1048)		(0.1016)	(0.0853)		(0.1121)
motorcycleyes	0.6104	0.4823	0.5883	0.5214	[0.2471, 0.8125]	0.5154
	(0.1783)		(0.1334)	(0.1702)		(0.1625)
caryes	0.2741	—	0.5805	0.5544	[-0.0336, 1.0132]	0.5470
	(0.2058)		(0.2378)	(0.2610)		(0.2896)

where $\mathcal{T} \subset (0,1)$ is a fixed compact set. Indeed, for each $\tau \in \mathcal{T}$ and each $k = 1, \ldots, K$, estimates can be obtained by applying the methods to the model (2.1) as

$$\tau - \text{quantile}(y \mid z, d) = d_k \alpha_{\tau k} + g_\tau(z) \quad \text{where} \quad g_\tau(z) := \tilde{g}_\tau(z) + \sum_{j \neq k} d_j \alpha_{\tau j}.$$

For each $\tau \in \mathcal{T}$, Step 1 and the conditional density function f_i, $i = 1, \ldots, n$, are the same for all $k = 1, \ldots, K$. However, Steps 2 and 3 adapt to each quantile index and each coefficient of interest. The uniform validity of ℓ_1-penalized methods for a continuum of problems (indexed by $\mathcal{T}$ in our case) has been established for quantile regression in [3] and for least squares in [11]. These works established that in a variety of settings we obtain the same rate of convergence as for a single index (provided $\mathcal{T}$ is a compact set with a dimension that does not grow).

The results obtained here lead directly to marginal confidence intervals that are valid for all $\alpha_{\tau k}$, $k = 1, \ldots, K$, $\tau \in \mathcal{T}$, namely

$$\lim_{n \to \infty} \sup_{\mathrm{P} \in \mathcal{P}_n} \sup_{t \in \mathbb{R}} \sup_{\tau \in \mathcal{T}} \max_{k=1,\ldots,K} |\mathrm{P}(\sigma_{n\tau k}^{-1} \sqrt{n}(\check{\alpha}_{\tau k} - \alpha_{\tau k}) \leqslant t) - \mathrm{P}(N(0,1) \leqslant t)| = 0$$

where $\sigma_{n\tau k}^2 = \tau(1-\tau)\bar{\mathrm{E}}[v_{\tau ki}^2]^{-1}$.

Furthermore, uniform confidence bands are possible by defining the critical value

$$c^*(1-\xi) = \inf \left\{ t : \mathrm{P}\left(\sup_{\tau \in \mathcal{T}, k=1,\ldots,K} |\mathbb{U}_n(\tau, k)| \leqslant t \mid \{d_i, z_i\}_{i=1}^n \right) \geqslant 1 - \xi \right\},$$

where the random variable $\mathbb{U}_n(\tau, k)$ is pivotal conditional on the data, namely

$$\mathbb{U}_n(\tau, k) := \frac{\{\tau(1-\tau)\bar{\mathrm{E}}[v_{\tau ki}^2]\}^{-1/2}}{\sqrt{n}} \sum_{i=1}^n (\tau - 1\{U_i \leqslant \tau\}) v_{\tau ki}$$

where U_i are i.i.d. uniform $(0,1)$ random variables, independent of $\{d_i, z_i\}_{i=1}^n$. Therefore $c^*(1-\xi)$ can be estimated since estimates of $v_{\tau ki}$ and $\sigma_{n\tau k}$, $\tau \in \mathcal{T}$ and $k = 1, \ldots, K$, are available. Uniform confidence bands can be defined as

$$[\check{\alpha}_{\tau k} - \sigma_{n\tau k} c^*(1-\xi)/\sqrt{n}, \check{\alpha}_{\tau k} + \sigma_{n\tau k} c^*(1-\xi)/\sqrt{n}] \quad \text{for} \quad \tau \in \mathcal{T}, \ k = 1, \ldots, K.$$

5.3. **Handling Approximately Sparse Functions.** As discussed in Remark 2.1, in order to handle approximately sparse models to represent g_τ in (2.3) an approximately orthogonality condition is assumed, namely

$$\bar{\mathrm{E}}[f_i v_i r_{g\tau i}] = o(n^{-1/2}). \tag{5.27}$$

In the literature such condition has been (implicitly) used before. For example, (5.27) holds if the function g_τ is exactly sparse linear combination of the covariates so that all the approximation errors $r_{g\tau i} = 0, i = 1, \ldots, n$. An alternative assumption in the literature that implies (5.27) is to have $\mathrm{E}[f_i d_i \mid z_i] = f_i\{x_i'\theta_\tau + r_{\theta\tau i}\}$, where θ_τ is sparse and $r_{\theta\tau i}$ is suitably small, which implies orthogonality to all functions of z_i since we have $\mathrm{E}[f_i v_i \mid z_i] = 0$.

The high-dimensional setting make the condition (5.27) less restrictive as p grows. Our discussion is based on the assumption that the function g_τ belongs to a well behaved class of functions. For example, when g_τ belongs to a Sobolev space $\mathcal{S}(\alpha, L)$ for some $\alpha \geqslant 1$ and $L > 0$ with respect to the basis

$\{x_j = P_j(z), j \geqslant 1\}$. As in [35], a Sobolev space of functions consists of functions $g(z) = \sum_{j=1}^{\infty} \theta_j P_j(z)$ whose Fourier coefficients θ satisfy

$$\theta \in \Theta(\alpha, L) = \left\{ \theta \in \ell^2(\mathbb{N}) : \ \sum_{j=1}^{\infty} |\theta_j| < \infty, \ \ \sum_{j=1}^{\infty} j^{2\alpha} \theta_j^2 \leqslant L^2 \ \right\}.$$

More generally, we can consider functions in a p-Rearranged Sobolev space $\mathcal{RS}(\alpha, p, L)$ which allow permutations in the first p components as in [12]. Formally, the class of functions $g(z) = \sum_{j=1}^{\infty} \theta_j P_j(z)$ such that

$$\theta \in \Theta^R(\alpha, p, L) = \left\{ \theta \in \ell^2(\mathbb{N}) : \sum_{j=1}^{\infty} |\theta_j| < \infty, \quad \begin{array}{l} \exists \text{ permutation } \Upsilon : \{1, \ldots, p\} \to \{1, \ldots, p\} \\ \sum_{j=1}^{p} j^{2\alpha} \theta_{\Upsilon(j)}^2 + \sum_{j=p+1}^{\infty} j^{2\alpha} \theta_j^2 \leqslant L^2 \end{array} \right\}.$$

It follows that $\mathcal{S}(\alpha, L) \subset \mathcal{RS}(\alpha, p, L)$ and p-Rearranged Sobolev space reduces substantially the dependence on the ordering of the basis.

Under mild conditions, it was shown in [12] that for functions in $\mathcal{RS}(\alpha, p, L)$ the rate-optimal choice for the size of the support of the oracle model obeys $s \lesssim n^{1/[2\alpha+1]}$. It follows that

$$\bar{\mathrm{E}}[r_{g\tau}^2]^{1/2} = \bar{\mathrm{E}}[\{\textstyle\sum_{j>s} \theta_{(j)} P_{(j)}(z_i)\}^2]^{1/2} \lesssim n^{-\alpha/\{1+2\alpha\}}.$$

which cannot guarantee converge to zero at a $\sqrt{n}$-rate to potentially imply (5.27). However, the relation (2.10) can exploit orthogonality with respect all p components of x_i, namely

$$\begin{aligned} |\bar{\mathrm{E}}[f_i v_i r_{g\tau i}]| \ &= |\bar{\mathrm{E}}[f_i v_i \{\textstyle\sum_{j=s+1}^{p} \theta_j P_j(z_i) + \sum_{j \geqslant p+1} \theta_j P_j(z_i)]| \\ &= |\textstyle\sum_{j \geqslant p+1} \bar{\mathrm{E}}[f_i v_i \theta_j P_j(z_i)]| \leqslant \sum_{j \geqslant p+1} |\theta_j| \{\bar{\mathrm{E}}[f_i^2 v_i^2] \mathrm{E}[P_j^2(z_i)]\}^{1/2} \\ &\leqslant \{\bar{\mathrm{E}}[f_i^2 v_i^2] \max_{j \geqslant p+1} \mathrm{E}[P_j^2(z_i)]\}^{1/2} (\textstyle\sum_{j \geqslant p+1} |\theta_j|^2 j^{2\alpha})^{1/2} (\sum_{j \geqslant p+1} j^{-2\alpha})^{1/2} = O(p^{-\alpha+1/2}). \end{aligned}$$

Therefore, condition (5.27) holds if $n = o(p^{2\alpha-1})$, in particular, for any $\alpha \geqslant 1$, $n = o(p)$ suffices.

5.4. **Minimax Efficiency.** In this section we make some connections to the (local) minimax efficiency analysis from the semiparametric efficiency analysis. In this section for the sake of exposition we assume that $(y_i, x_i, d_i)_{i=1}^n$ are i.i.d., sparse models, $r_{\theta\tau i} = r_{g\tau i} = 0$, $i = 1, \ldots, n$, and the median case ($\tau = .5$). [25] derives an efficient score function for the partially linear median regression model:

$$S_i = 2\varphi_\tau(y_i, d_i\alpha_\tau + x_i'\beta_\tau) f_i [d_i - m_\tau^*(z)],$$

where $m_\tau^*(z_i)$ is given by

$$m_\tau^*(z_i) = \frac{\mathrm{E}[f_i^2 d_i | z_i]}{\mathrm{E}[f_i^2 | z_i]}.$$

Using the assumption $m_\tau^*(z_i) = x_i'\theta_\tau^*$, where $\|\theta_\tau^*\|_0 \leqslant s \ll n$ is sparse, we have that

$$S_i = 2\varphi_\tau(y_i, d_i\alpha_\tau + x_i'\beta_\tau) v_i^*,$$

where $v_i^* = f_i d_i - f_i m_\tau^*(z_i)$ would correspond to v_i in (2.7). It follows that the estimator based on the instrument v_i^* is actually efficient in the minimax sense (see Theorem 18.4 in [23]), and inference about α_τ based on this estimator provides best minimax power against local alternatives (see Theorem 18.12 in [23]).

The claim above is formal as long as, given a law Q_n, the least favorable submodels are permitted as deviations that lie within the overall model. Specifically, given a law Q_n, we shall need to allow for a

certain neighborhood $\mathcal{Q}_n^\delta$ of Q_n such that $Q_n \in \mathcal{Q}_n^\delta \subset \mathcal{Q}_n$, where the overall model $\mathcal{Q}_n$ 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 $t = (t_1, t_2)$:

$$y_i = d_i(\alpha_\tau + t_1) + x_i'(\beta_\tau + t_2\theta_\tau^*) + \epsilon_i, \quad \|t\| \leqslant \delta, \tag{5.28}$$

$$f_i d_i = f_i x_i'\theta_\tau^* + v_i^*, \quad \mathrm{E}[f_i v_i^* | x_i] = 0, \tag{5.29}$$

where $\|\beta_\tau\|_0 \vee \|\theta_\tau^*\|_0 \leqslant s/2$ and conditions as in Section 2 hold. The case with $t = 0$ generates the model Q_n; by varying t within δ-ball, we generate models $\mathcal{Q}_n^\delta$, containing the least favorable deviations. By [25], 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 $J = \mathrm{E}[S_i^2]$. Since our model $\mathcal{Q}_n$ contains $\mathcal{Q}_n^\delta$, 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 $\mathcal{Q}_n^t$ with $t = (O(1/\sqrt{n}), o(1))$, their first order asymptotics do not change, as a consequence of Theorems in Section 2. (Though our theorems actually prove more than this.)

Acknowledgements

We would like to specially thank Roger Koenker for providing the data for the empirical example and for many insightful discussions on inference. We would also like to thank the participants of the December 2012 Luminy conference on Nonparametric and high-dimensional statistics, the November 2012 Oberwolfach workshop on Frontiers in Quantile Regression, the August 2012 8th World Congress in Probability and Statistics, and a seminar at the University of Michigan.

Appendix A. Analysis under High-Level Conditions

This section contains the main tools used in establishing the main inferential results. The high-level conditions here are intended to be applicable in a variety of settings and they are implied by the regularities conditions provided in the previous sections. The results provided here are of independent interest (e.g. properties of Lasso under estimated weights). We establish the inferential results (2.17) and (2.20) in Section A.3 under high level conditions. To verify these high-level conditions we need rates of convergence for the estimated instruments $\widehat{v}$ and the estimated confounding function $\widehat{g}_\tau(z) = x'\widehat{\beta}_\tau$ which are established in sections A.2 and A.1 respectively. The main design condition relies on the restricted eigenvalue proposed in [13], namely for $\tilde{x}_i = [d_i, x_i']'$

$$\kappa_{\mathbf{c}} = \inf_{\|\delta_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1} \|\tilde{x}_i'\delta\|_{2,n} / \|\delta_T\| \tag{A.30}$$

where $\mathbf{c} = (c+1)/(c-1)$ for the slack constant $c > 1$, see [13]. It is well known that Condition SE implies that $\kappa_{\mathbf{c}}$ is bounded away from zero if $\mathbf{c}$ is bounded, see [13].

A.1. **ℓ_1-Penalized Quantile Regression.** In this section for a quantile index $u \in (0, 1)$, we consider the equation

$$\tilde{y}_i = \tilde{x}_i'\eta_u + r_{ui} + \epsilon_i, \quad u\text{-quantile of } (\epsilon_i \mid \tilde{x}_i, r_{ui}) = 0 \tag{A.31}$$

where we observe $\{(\tilde{y}_i, \tilde{x}_i) : i = 1, \ldots, n\}$, which are independent across i. To estimate η_u we consider the ℓ_1-penalized u-quantile regression estimate

$$\widehat{\eta}_u \in \arg\min_{\eta} \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta)] + \frac{\lambda_u}{n}\|\eta\|_1$$

and the associated post-model selection estimate

$$\widetilde{\eta}_u \in \arg\min_{\eta} \{ \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta)] \ : \ \eta_j = 0 \ \text{ if } \ \widehat{\eta}_{uj} = 0\}. \tag{A.32}$$

As established in [3] for sparse models and in [19] for approximately sparse models, under the event that

$$\frac{\lambda_u}{n} \geqslant c\|\mathbb{E}_n[(u - 1\{\tilde{y}_i \leqslant \tilde{x}_i'\eta_u + r_{ui}\})\tilde{x}_i]\|_\infty \tag{A.33}$$

the estimator above achieves good theoretical guarantees under mild design conditions. Although η_u is unknown, we can set λ_u so that the event in (A.33) holds with high probability. In particular, the pivotal rule proposed in [3] and generalized in [19] proposes to set $\lambda_u := cn\Lambda_u(1 - \gamma \mid \tilde{x})$ for $c > 1$ where

$$\Lambda_u(1 - \gamma \mid \tilde{x}) = (1 - \gamma) - \text{quantile of } \|\mathbb{E}_n[(u - 1\{U_i \leqslant u\})\tilde{x}_i]\|_\infty \tag{A.34}$$

where $U_i \sim U(0, 1)$ are independent random variables conditional on $\tilde{x}_i$, $i = 1, \ldots, n$. This quantity can be easily approximated via simulations. Below we summarize the high level conditions we require.

Condition PQR. Let $T_u = \text{support}(\eta_u)$ and normalize $\mathbb{E}_n[\tilde{x}_{ij}^2] = 1$, $j = 1, \ldots, p$. Assume that for some $s \geqslant 1$, $\|\eta_u\|_0 \leqslant s$, $\|r_{ui}\|_{2,n} \leqslant C\sqrt{s/n}$. Further, the conditional distribution function of ϵ_i is absolutely continuous with continuously differentiable density $f_\epsilon(\cdot \mid d_i, z_i)$ such that $0 < \underline{f} \leqslant f_i \leqslant \sup_\epsilon f_{\epsilon_i|d_i,z_i}(\epsilon \mid d_i, z_i) \leqslant \bar{f}$, $\sup_\epsilon f'_{\epsilon_i|d_i,z_i}(\epsilon \mid d_i, z_i) < \bar{f}'$ for fixed constants $\underline{f}$, $\bar{f}$ and $\bar{f}'$.

Condition PQR is implied by Condition AS. The conditions on the approximation error and near orthogonality conditions follows from choosing a model η_u that optimally balance the bias/variance trade-off. The assumption on the conditional density is standard in the quantile regression literature even with fixed p case developed in [22] or the case of p increasing slower than n studied in [5].

Next we present bounds on the prediction norm of the ℓ_1-penalized quantile regression estimator.

Lemma 1 (Estimation Error of ℓ_1-Penalized Quantile Regression)**.** *Under Condition PQR, setting $\lambda_u \geqslant cn\Lambda_u(1 - \gamma \mid \tilde{x})$, we have with probability $1 - 4\gamma$ for n large enough*

$$\|\tilde{x}_i'(\widehat{\eta}_u - \eta_u)\|_{2,n} \lesssim N := \frac{\lambda_u\sqrt{s}}{n\kappa_{2\mathbf{c}}} + \frac{1}{\kappa_{2\mathbf{c}}}\sqrt{\frac{s\log(p/\gamma)}{n}}$$

provided that for $A_u := \Delta_{2\mathbf{c}} \cup \{v : \|\tilde{x}_i'v\|_{2,n} = N, \|v\|_1 \leqslant 8C\mathbf{c}s\log(p/\gamma)/\lambda_u\}$, we have

$$\sup_{\bar{\delta} \in A_u} \frac{\mathbb{E}_n[|r_{ui}||\tilde{x}_i'\bar{\delta}|^2]}{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^2]} + N \sup_{\bar{\delta} \in A_u} \frac{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^3]}{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^2]^{3/2}} \to 0.$$

Lemma 1 establishes the rate of convergence in the prediction norm for the ℓ_1-penalized quantile regression estimator. Exact constants are derived in the proof. The extra growth condition required for

identification is mild. For instance we typically have $\lambda_u \sim \sqrt{\log(n \vee p)/n}$ and for many designs of interest we have $\inf_{\delta \in \Delta_{\mathbf{c}}} \|\tilde{x}_i'\delta\|_{2,n}^3 / \mathbb{E}_n[|\tilde{x}_i'\delta|^3]$ bounded away from zero (see [3]). For more general designs we have

$$\inf_{\delta \in A_u} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \geqslant \inf_{\delta \in A_u} \frac{\|\tilde{x}_i'\delta\|_{2,n}}{\|\delta\|_1 \max_{i \leqslant n} \|\tilde{x}_i\|_\infty} \geqslant \frac{\kappa_{2\mathbf{c}}}{\sqrt{s}(1+\mathbf{c}) \max_{i \leqslant n} \|\tilde{x}_i\|_\infty} \wedge \frac{\lambda_u N}{8C\mathbf{c}s \log(p/\gamma) \max_{i \leqslant n} \|\tilde{x}_i\|_\infty}.$$

Lemma 2 (Estimation Error of Post-ℓ_1-Penalized Quantile Regression)**.** *Assume Condition PQR holds, and that the Post-ℓ_1-penalized quantile regression is based on an arbitrary vector $\widehat{\eta}_u$. Let $\bar{r}_u \geqslant \|r_{ui}\|_{2,n}$, $\widehat{s}_u \geqslant |\mathrm{support}(\widehat{\eta}_u)|$ and $\widehat{Q} \geqslant \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\widehat{\eta}_u)] - \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u))]$ hold with probability $1-\gamma$. Then we have for n large enough, with probability $1 - \gamma - \epsilon - o(1)$*

$$\|\tilde{x}_i'(\widetilde{\eta}_u - \eta_u)\|_{2,n} \lesssim \widetilde{N} := \sqrt{\frac{(\widehat{s}_u + s)\log(p/\varepsilon)}{n\phi_{\min}(\widehat{s}_u + s)}} + \bar{f}\bar{r}_u + \widehat{Q}^{1/2}$$

provided that

$$\sup_{\|\bar{\delta}\|_0 \leqslant \widehat{s}_u + s} \frac{\mathbb{E}_n[|r_{ui}||\tilde{x}_i'\bar{\delta}|^2]}{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^2]} + \widetilde{N} \sup_{\|\bar{\delta}\|_0 \leqslant \widehat{s}_u + s} \frac{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^3]}{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^2]^{3/2}} \to 0.$$

Lemma 2 provides the rate of convergence in the prediction norm for the post model selection estimator despite of possible imperfect model selection. In the current nonparametric setting it is unlikely for the coefficients to exhibit a large separation from zero. The rates rely on the overall quality of the selected model by ℓ_1-penalized quantile regression and the overall number of components $\widehat{s}_u$. Once again the extra growth condition required for identification is mild. For more general designs we have

$$\inf_{\|\delta\|_0 \leqslant \widehat{s}_u + s} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \geqslant \inf_{\|\delta\|_0 \leqslant \widehat{s}_u + s} \frac{\|\tilde{x}_i'\delta\|_{2,n}}{\|\delta\|_1 \max_{i \leqslant n} \|\tilde{x}_i\|_\infty} \geqslant \frac{\sqrt{\phi_{\min}(\widehat{s}_u + s)}}{\sqrt{\widehat{s}_u + s} \max_{i \leqslant n} \|\tilde{x}_i\|_\infty}.$$

A.2. **Lasso with Estimated Weights.** In this section we consider the equation

$$f_i d_i = f_i m_\tau(z_i) + v_i = f_i x_i'\theta_\tau + f_i r_{\theta\tau i} + v_i, \quad \mathrm{E}[f_i v_i x_i] = 0 \tag{A.35}$$

where we observe $\{(d_i, z_i, x_i = P(z_i)) : i = 1, \ldots, n\}$, which are independent across i. We do not observe $\{f_i = f_\tau(d_i, z_i)\}_{i=1}^n$ directly, but we assume that estimates $\{\widehat{f}_i\}_{i=1}^n$ are available. Also, we have that $T_{\theta\tau} = \mathrm{support}(\theta_\tau)$ is unknown but a sparsity condition holds, namely $|T_{\theta\tau}| \leqslant s$. To estimate $\theta_{\theta\tau}$ and v_i, we compute

$$\widehat{\theta}_\tau \in \arg\min_\theta \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta)^2] + \frac{\lambda}{n}\|\widehat{\Gamma}_\tau\theta\|_1 \ \text{ and set } \ \widehat{v}_i = \widehat{f}_i(d_i - x_i'\widehat{\theta}_\tau), \quad i = 1, \ldots, n, \tag{A.36}$$

where λ and $\widehat{\Gamma}_\tau$ are the associated penalty level and loadings specified below. The new difficulty is to account for the impact of estimated weights $\widehat{f}_i$. Although this impact on the estimation of θ_τ is minor, the estimated weights impact estimates of v_i can be more substantial.

We will establish bounds on the penalty parameter λ so that with high probability the following regularization event occurs

$$\frac{\lambda}{n} \geqslant 2c\|\widehat{\Gamma}_\tau^{-1}\mathbb{E}_n[f_i x_i v_i]\|_\infty. \tag{A.37}$$

As discussed in [13, 4, 9], the event above allows to exploit the restricted set condition $\|\widehat{\theta}_{\tau T_{\theta\tau}^c}\|_1 \leqslant \tilde{\mathbf{c}}\|\widehat{\theta}_{\tau T_{\theta\tau}} - \theta_\tau\|_1$ for some $\tilde{\mathbf{c}} > 1$. Thus rates of convergence for $\widehat{\theta}_\tau$ and $\widehat{v}_i$ defined on (A.36) can be established based on the restricted eigenvalue $\kappa_{\tilde{\mathbf{c}}}$ defined in (A.30) with $\tilde{x}_i = x_i$.

However, the estimation error in the estimate $\widehat{f}_i$ of f_i could slow the rates of convergence. The following are sufficient high-level conditions.

Condition WL. *For the model (A.35), normalize* $\mathbb{E}_n[x_{ij}^2] = 1$, $j = 1, \ldots, p$, *and suppose that:*
(i) for $s \geqslant 1$ *we have* $\|\theta_\tau\|_0 \leqslant s$, $\mathbb{E}_n[r_{\theta\tau i}^2] \leqslant Cs/n$, $\Phi^{-1}(1-\gamma/2p) \leqslant \delta_n n^{1/6}$,
(ii) $0 < \underline{f} \leqslant f_i \leqslant \bar{f}$ *uniformly in* n, *and* $0 < \underline{c} \leqslant \mathrm{E}[v_i^2 \mid x_i] \leqslant \bar{c} < \infty$, *a.s.,* $\max_{j \leqslant p} \frac{\{\bar{\mathrm{E}}[|f_i x_{ij} v_i|^3]\}^{1/3}}{\{\bar{\mathrm{E}}[|f_i x_{ij} v_i|^2]\}^{1/2}} \leqslant C$,
(iii) with probability $1 - \Delta_n$ *we have* $\max_{i \leqslant n} \|x_i\|_\infty \leqslant K_x$,

$$\max_{j \leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[f_i^2 x_{ij}^2 v_i^2]| \leqslant \delta_n, \quad \max_{j \leqslant p} \mathbb{E}_n[(\widehat{f}_i - f_i)^2 x_{ij}^2 v_i^2] \leqslant \delta_n, \quad \mathbb{E}_n[\widehat{f}_i^2 r_{\theta\tau i}^2] \leqslant c_r^2, \quad \mathbb{E}_n\left[\tfrac{(\widehat{f}_i^2 - f_i^2)^2}{f_i^2} v_i^2\right] \leqslant c_f^2.$$

(iv) $\ell\widehat{\Gamma}_{\tau 0} \leqslant \widehat{\Gamma}_\tau \leqslant u\widehat{\Gamma}_{\tau 0}$, *where* $\widehat{\Gamma}_{\tau 0 jj} = \{\mathbb{E}_n[\widehat{f}_i^2 x_{ij}^2 v_i^2]\}^{1/2}$, $1 - \delta_n \leqslant \ell \leqslant u \leqslant C$ *with probability* $1 - \Delta_n$.

Comment A.1. Condition WL(i) is a standard condition on the approximation error that yields the optimal bias variance trade-off (see [4]) and imposes a growth restriction on p relative to n, in particular $\log p = o(n^{1/3})$. Condition WL(ii) imposes conditions on the conditional density function and mild moment conditions which are standard in quantile regression models even with fixed dimensions, see [22]. Condition WL(iii) requires high-level rates of convergence for the estimate $\widehat{f}_i$. Several primitive moment conditions imply first requirement in Condition WL(iii). These conditions allow the use of self-normalized moderate deviation theory to control heteroscedastic non-Gaussian errors similarly to [2] where there are no estimated weights. Condition WL(iv) corresponds to the asymptotically valid penalty loading in [2] which is satisfied by the proposed choice $\widehat{\Gamma}_\tau$ in (2.12). □

Next we present results on the performance of the estimators generated by Lasso with estimated weights.

Lemma 3 (Rates of Convergence for Lasso)**.** *Under Condition WL and setting* $\lambda \geqslant 2c'\sqrt{n}\Phi^{-1}(1-\gamma/2p)$ *for* $c' > c > 1$, *we have for* n *large enough with probability* $1 - \gamma - o(1)$

$$\|\widehat{f}_i x_i'(\widehat{\theta}_\tau - \theta_\tau)\|_{2,n} \leqslant 2\{c_f + c_r\} + \frac{\lambda\sqrt{s}}{n\widehat{\kappa}_{\tilde{\mathbf{c}}}}\left(u + \frac{1}{c}\right)$$
$$\|\widehat{\theta}_\tau - \theta_\tau\|_1 \leqslant 2\frac{\sqrt{s}\{c_f + c_r\}}{\widehat{\kappa}_{2\tilde{\mathbf{c}}}} + \frac{\lambda s}{n\widehat{\kappa}_{\tilde{\mathbf{c}}}\widehat{\kappa}_{2\tilde{\mathbf{c}}}}\left(u + \frac{1}{c}\right) + \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\frac{2c\|\widehat{\Gamma}_{\tau 0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}\{c_f + c_r\}^2$$

where $\tilde{\mathbf{c}} = \|\widehat{\Gamma}_{\tau 0}^{-1}\|_\infty \|\widehat{\Gamma}_{\tau 0}\|_\infty (uc + 1)/(\ell c - 1)$

Lemma 3 above establishes the rate of convergence for Lasso with estimated weights. This automatically leads to bounds on the estimated instrument $\widehat{v}_i$ obtained with Lasso through the identity

$$\widehat{v}_i - v_i = (\widehat{f}_i - f_i)\frac{v_i}{f_i} + \widehat{f}_i x_i'(\theta_\tau - \widehat{\theta}_\tau) + \widehat{f}_i r_{\theta\tau i}. \tag{A.38}$$

The Post-Lasso estimator applies the least squares estimator to the model selected by the Lasso estimator (A.36),

$$\widetilde{\theta}_\tau \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta)^2] \ : \ \theta_j = 0, \text{ if } \widehat{\theta}_{\tau j} = 0 \right\}, \quad \text{set } \tilde{v}_i = \widehat{f}_i(d_i - x_i'\widetilde{\theta}_\tau).$$

It aims to remove the bias towards zero induced by the ℓ_1-penalty function which is used to select components. Sparsity properties of the Lasso estimator $\widehat{\theta}_\tau$ under estimated weights follows similarly to the standard Lasso analysis derived in [2]. By combining such sparsity properties and the rates in the

prediction norm we can establish rates for the post-model selection estimator under estimated weights. The following result summarizes the properties of the Post-Lasso estimator.

Lemma 4 (Model Selection Properties of Lasso and Properties of Post-Lasso)**.** *Suppose that Condition WL holds, and* $\kappa' \leqslant \phi_{\min}(\{s + \frac{n^2}{\lambda^2}\{c_f^2 + c_r^2\}\}/\delta_n) \leqslant \phi_{\max}(\{s + \frac{n^2}{\lambda^2}\{c_f^2 + c_r^2\}\}/\delta_n) \leqslant \kappa''$ *for some positive and bounded constants* κ', κ''*. Then the data-dependent model* $\widehat{T}_{\theta\tau}$ *selected by the Lasso estimator with* $\lambda \geqslant 2c'\sqrt{n}\Phi^{-1}(1-\gamma/2p)$ *for* $c' > c > 1$*, satisfies with probability* $1 - \gamma - o(1)$*:*

$$\|\widetilde{\theta}_\tau\|_0 = |\widehat{T}_{\theta\tau}| \lesssim s + \frac{n^2}{\lambda^2}\{c_f^2 + c_r^2\} \tag{A.39}$$

Moreover, the corresponding Post-Lasso estimator obeys

$$\|x_i'(\widetilde{\theta}_\tau - \theta_\tau)\|_{2,n} \lesssim_P c_f + c_r + \sqrt{\frac{|\widehat{T}_{\theta\tau}| \log(p \vee n)}{n}} + \frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}}.$$

A.3. **Instrumental Quantile Regression with Estimated Data.** Next we turn to analyze the instrumental quantile regression discussed in Section 2. Condition IQR below suffices to make the impact of the estimation of instruments negligible to the first order asymptotics of the estimator $\check{\alpha}_\tau$. Primitive conditions that imply Condition IQR are provided and discussed in the main text.

Let $(d, z) \in \mathcal{D} \times \mathcal{Z}$. In this section for $\tilde{h} = (\tilde{g}, \tilde{\iota})$, where $\tilde{g}$ is a function of variable z, and the instrument $\tilde{\iota}$ is a function that maps $(d, x) \mapsto \tilde{\iota}(d, z)$ we write

$$\psi_{\tilde{\alpha},\tilde{h}}(y_i, d_i, z_i) = \psi_{\tilde{\alpha},\tilde{g},\tilde{\iota}}(y_i, d_i, z_i) = (\tau - 1\{y_i \leqslant \tilde{g}(z_i) + d_i\alpha\})\tilde{\iota}(d_i, z_i) = (\tau - 1\{y_i \leqslant \tilde{g}_i + d_i\alpha\})\tilde{\iota}_i.$$

We assume that the estimated functions $\widehat{g}$ and $\widehat{\iota}$ satisfy the following condition.

Condition IQR. *Let* $\{(y_i, d_i, z_i) : i = 1, \ldots, n\}$ be independent observations satisfying (2.1). Suppose that there are positive constants $0 < c \leqslant C < \infty$ such that:
(i) $f_{y_i|d_i,z_i}(y \mid d_i, z_i) \leqslant \bar{f}$, $f'_{y_i|d_i,z_i}(y \mid d_i, z_i) \leqslant \bar{f}'$; $c \leqslant |\bar{\mathrm{E}}[f_i d_i \iota_{0i}]|$, and $\max_{i \leqslant n}\{\mathrm{E}[\iota_{0i}^4]\}^{1/2} \vee \{\mathrm{E}[d_i^4]\}^{1/2} \leqslant C$;
(ii) $\{\alpha : |\alpha - \alpha_\tau| \leqslant n^{-1/2}/\delta_n\} \subset \mathcal{A}_\tau$, where $\mathcal{A}_\tau$ is a (possibly random) compact interval;
(iii) For some sequences $\delta_n \to 0$ and $\Delta_n \to 0$ with probability at least $1 - \Delta_n$ the estimated quantities $\widehat{h} = (\widehat{g}, \widehat{\iota})$ satisfy

$$\begin{aligned} &\max_{i \leqslant n}\{1 + |\iota_{0i}| + |\widehat{\iota}_i - \iota_{0i}|\}^{1/2}\|g_{\tau i} - \widehat{g}_i\|_{2,n} \leqslant \delta_n n^{-1/4}, \\ &\|\widehat{\iota}_i - \iota_{0i}\|_{2,n} \leqslant \delta_n, \quad \|g_{\tau i} - \widehat{g}_i\|_{2,n} \cdot \|\widehat{\iota}_i - \iota_{0i}\|_{2,n} \leqslant \delta_n n^{-1/2}, \end{aligned} \tag{A.40}$$

$$|\mathbb{E}_n[f_i \iota_{0i}\{\widehat{g}_i - g_{\tau i}\}]| \leqslant \delta_n n^{-1/2}, \tag{A.41}$$

$$\sup_{\alpha \in \mathcal{A}_\tau} \left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\psi_{\alpha,\widehat{h}}(y_i, d_i, z_i) - \psi_{\alpha,h_0}(y_i, d_i, z_i)\right]\right| \leqslant \delta_n\, n^{-1/2} \tag{A.42}$$

$$|\check{\alpha}_\tau - \alpha_\tau| \leqslant \delta_n \quad \text{and} \quad \mathbb{E}_n[\psi_{\check{\alpha}_\tau,\widehat{h}}(y_i, d_i, z_i)]| \leqslant \delta_n\, n^{-1/2} \tag{A.43}$$

(iv) $\|\widehat{\iota}_i - \iota_{0i}\|_{2,n} \leqslant \delta_n$ and $\|1\{|\epsilon_i| \leqslant |d_i(\alpha_\tau - \check{\alpha}_\tau) + g_{\tau i} - \widehat{g}_i|\}\|_{2,n} \leqslant \delta_n^2$.

Lemma 5. *Under Condition IQR(i,ii,iii) we have*

$$\bar{\sigma}_n^{-1}\sqrt{n}(\check{\alpha}_\tau - \alpha_\tau) = \mathbb{U}_n(\tau) + o_P(1), \quad \mathbb{U}_n(\tau) \rightsquigarrow N(0, 1)$$

where $\bar{\sigma}_n^2 = \bar{\mathrm{E}}[f_i d_i \iota_{0i}]^{-1}\bar{\mathrm{E}}[\tau(1-\tau)\iota_{0i}^2]\bar{\mathrm{E}}[f_i d_i \iota_{0i}]^{-1}$ *and*

$$\mathbb{U}_n(\tau) = \{\bar{\mathrm{E}}[\psi_{\alpha_\tau,h_0}^2(y_i, d_i, z_i)]\}^{-1/2}\sqrt{n}\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i, d_i, z_i)].$$

Moreover, if additionally IQR(iv) holds we have

$$nL_n(\alpha_\tau) = \mathbb{U}_n(\tau)^2 + o_P(1), \quad \mathbb{U}_n(\tau)^2 \rightsquigarrow \chi^2(1)$$

and the variance estimator is consistent, namely

$$\mathbb{E}_n[\widehat{f}_i d_i \widehat{\iota}_i]^{-1} \mathbb{E}_n[(\tau - 1\{y_i \leqslant \widehat{g}_i + d_i \check{\alpha}_\tau\})^2 \widehat{\iota}_i^2] \mathbb{E}_n[\widehat{f}_i d_i \widehat{\iota}_i]^{-1} \to_P \bar{\mathrm{E}}[f_i d_i \iota_{0i}]^{-1} \bar{\mathrm{E}}[\tau(1-\tau)\iota_{0i}^2] \bar{\mathrm{E}}[f_i d_i \iota_{0i}]^{-1}.$$

Appendix B. Results for Section 3

Proof of Theorem 1. We will verify Condition IQR and the result follows by Lemma 5 noting that $1\{y_i \leqslant \alpha_\tau d_i + g_\tau(z_i)\} = 1\{U_i \leqslant \tau\}$ for some uniform $(0,1)$ random variable (independent of $\{d_i, z_i\}$) by the definition of the conditional quantile function.

Condition IQR(i) is assumed. Condition SE implies that $\kappa_{\mathbf{c}}$ is bounded away from zero for n sufficiently large. Step 1 relies on Post-ℓ_1-qr. By the truncation we have $\widehat{s}_\tau = \|\widetilde{\beta}_\tau\|_0 \leqslant Cs$ for any $C \geqslant 2$. Thus, by Condition SE, for large enough n we have that $\phi_{\min}(\widehat{s}_\tau + s)$ is bounded away from zero with probability $1 - \Delta_n$ since $\widehat{s}_\tau + s \leqslant \ell_n s$. Moreover, Condition PQR is implied by Condition AS. Lemma 6 ensures that $\|\widehat{\beta}_\tau^{(2s)} - \beta_\tau\|_1 \leqslant 2\|\widehat{\beta}_\tau - \beta_\tau\|_1$ and $\|x_i'(\widehat{\beta}_\tau^{(2s)} - \beta_\tau)\|_{2,n} \leqslant \|x_i'(\widehat{\beta}_\tau - \beta_\tau)\|_{2,n} + \sqrt{\phi_{\max}(s)/s}\|\widehat{\beta}_\tau - \beta_\tau\|_1$ since $\phi_{\max}(k)/k$ is decreasing in k. Therefore, by Lemma 2 we have $\|x_i'(\widetilde{\beta}_\tau - \beta_\tau)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n}$ provided the side conditions required in Lemmas 1 and 2. To verify those side conditions for Lemma 2 let $\tilde{x}_i = (d_i, x_i')'$ and $\delta = (\delta_d, \delta_x')'$. By Condition SE and $\mathbb{E}_n[|d_i|^3] \lesssim_P \bar{\mathrm{E}}[|d_i|^3] \leqslant C$, we have

$$\begin{aligned}
\inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} &\geqslant \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\{\phi_{\min}(s+Cs)\}^{3/2}\|\delta\|^3}{4\mathbb{E}_n[|x_i'\delta_x|^3] + 4|\delta_d|^3\mathbb{E}_n[|d_i|^3]} \\
&\geqslant \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\{\phi_{\min}(s+Cs)\}^{3/2}\|\delta\|^3}{4K_x\|\delta_x\|_1\phi_{\max}(s+Cs)\|\delta_x\|^2 + 4\|\delta\|^3\mathbb{E}_n[|d_i|^3]} \\
&\geqslant \frac{\{\phi_{\min}(s+Cs)\}^{3/2}}{4K_x\sqrt{s+Cs}\phi_{\max}(s+Cs) + 4\mathbb{E}_n[|d_i|^3]} \gtrsim_P \frac{1}{K_x\sqrt{s}}.
\end{aligned}$$

The relation above and the conditions $K_x^2 s^2 \log^2(p \vee n) \leqslant \delta_n n$ and $\lambda_\tau \lesssim \sqrt{n\log(p \vee n)}$ yields

$$\frac{n\sqrt{\phi_{\min}(s+Cs)}}{\lambda_\tau\sqrt{s} + \sqrt{sn\log(p\vee n)}} \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \gtrsim_P \frac{\sqrt{n}}{K_x s\log(p\vee n)} \to \infty.$$

Moreover, noting that $\|\delta_x\| \vee |\delta_d| \leqslant \|\delta\|$, we have

$$\begin{aligned}
\sup_{\|\delta\|_0 \leqslant s+Cs} \frac{\mathbb{E}_n[|r_{g\tau i}|\ |\tilde{x}_i'\delta|^2]}{\|\tilde{x}_i'\delta\|_{2,n}^2} &\leqslant 2 \sup_{\|\delta\|_0 \leqslant s+Cs} \frac{\mathbb{E}_n[|r_{g\tau i}|\ |x_i'\delta_x|^2] + \mathbb{E}_n[|r_{g\tau i}|d_i^2\delta_d^2]}{\phi_{\min}(s+Cs)\|\delta\|^2} \\
&\leqslant 2 \sup_{\|\delta\|_0 \leqslant s+Cs} \frac{\|r_{g\tau i}\|_{2,n}\sqrt{\phi_{\max}(s+Cs)}\|\delta_x\|K_x\|\delta_x\|_1}{\phi_{\min}(s+Cs)\|\delta_x\|^2} + \frac{\|r_{g\tau i}\|_{2,n}\|d_i^2\|_{2,n}}{\phi_{\min}(s+Cs)} \\
&\leqslant C\sqrt{\tfrac{s}{n}}\frac{K_x\sqrt{s+Cs}}{\phi_{\min}(s+Cs)} \to 0.
\end{aligned}$$

The verification of the side condition for Lemma 1 follows similarly.

Step 2 relies on Post-Lasso. Condition WL(i) and (ii) are implied by Conditions AS. Indeed, the moment conditions in AS imply the first part, the second part $\Phi^{-1}(1 - \gamma/2p) \leqslant \delta_n n^{1/3}$ is implied by $\log(1/\gamma) \lesssim \log(p \vee n)$ and $\log^3 p \leqslant \delta_n n$. Next we establish Condition WL(iii) under known conditional density. The first condition is implied by Lemma 8 under the moment conditions and the growth condition $K_x^4 \log p \leqslant \delta_n n$ and $f_i \leqslant \bar{f}$. Since $\widehat{f}_i = f_i$ the other requirements in WL(iii) follows.

Next we establish Condition WL(iv). Note that

$$\max_{j\leqslant p}|\mathbb{E}_n[x_{ij}^2(f_id_i)^2]-\bar{\mathrm{E}}[x_{ij}^2(f_id_i)^2]|\leqslant\max_{j\leqslant p}|(\mathbb{E}_n-\bar{\mathrm{E}})[x_{ij}^2(f_id_i)^2]|\lesssim_P\delta_n\quad\text{and}$$
$$\max_{j\leqslant p}|\mathbb{E}_n[f_i^2x_{ij}^2v_i^2]-\bar{\mathrm{E}}[f_i^2x_{ij}^2v_i^2]|\leqslant\max_{j\leqslant p}|(\mathbb{E}_n-\bar{\mathrm{E}})[f_i^2x_{ij}^2v_i^2]|\lesssim_P\delta_n$$

by Lemma 8 because $\bar{\mathrm{E}}[f_i^2x_{ij}^2v_i^2]$ is bounded away from zero and from above. Thus $\widehat{\Gamma}_{\tau 0jj}$ is bounded away from zero and from above with probability $1-o(1)$. Next note that

$$\begin{aligned}(\max_{i\leqslant n}f_i^2)\bar{\mathrm{E}}[x_{ij}^2(f_id_i)^2]&\leqslant\overline{f}^4\max_{j\leqslant p}\bar{\mathrm{E}}[x_{ij}^2d_i^2]\leqslant C,\quad\text{and}\\(\max_{i\leqslant n}f_i^2)\bar{\mathrm{E}}[x_{ij}^2(f_id_i)^2]&=(\max_{i\leqslant n}f_i^2)\bar{\mathrm{E}}[x_{ij}^2(v_i^2+2m_{\tau i}(z_i)f_iv_i+f_i^2m_{\tau i}^2(z_i))]\\&=(\max_{i\leqslant n}f_i^2)\bar{\mathrm{E}}[x_{ij}^2v_i^2]+(\max_{i\leqslant n}f_i^2)\bar{\mathrm{E}}[x_{ij}^2f_i^2m_{\tau i}^2(z_i)]\geqslant\bar{\mathrm{E}}[f_i^2x_{ij}^2v_i^2].\end{aligned}$$

Therefore, the initial penalty loadings $\{(\max_{i\leqslant n}f_i^2)\mathbb{E}_n[x_{ij}^2(f_id_i)^2]\}_{j=1,\ldots,p}$ satisfy Condition WL(iv) with $\ell\to 1$ and $u\leqslant C$. By Lemma 4 and the growth conditions we have that the penalty loadings $\widehat{\Gamma}_{\tau jj}$ using $\widehat{v}_i$ also satisfy $\widehat{\Gamma}_{\tau 0jj}-\delta_n\leqslant\widehat{\Gamma}_{\tau jj}\leqslant u\widehat{\Gamma}_{\tau 0jj}$. Thus, by Lemma 4 we have $\|x_i'(\widetilde{\theta}_\tau-\theta_\tau)\|_{2,n}\lesssim_P\sqrt{s\log(n\vee p)/n}$ and $\|\widetilde{\theta}_\tau\|_0\lesssim_P s$.

Step 3 relies on instrumental quantile regression. Condition IQR(iii) relation (A.40) follows by the rates for $\widetilde{\beta}_\tau$ and $\widetilde{\theta}_\tau$ and the growth condition $K_x^2s^2\log^2(p\vee n)\leqslant\delta_nn$. To show the other relations in Condition IQR(iii) we will consider the class of functions

$$\mathcal{F}=\{1\{y_i\leqslant x_i'\beta+d_i\alpha\}-1\{y_i\leqslant x_i'\beta_\tau+d_i\alpha\}:\|\beta\|_0\leqslant Cs,\ \|\beta-\beta_\tau\|\leqslant C\sqrt{s\log(p\vee n)/n}\},$$

$$\text{and}\quad\mathcal{G}=\{\delta:\|x_i'\delta\|_{2,n}\leqslant C\sqrt{s\log(p\vee n)/n},\|\delta\|_0\leqslant Cs\}.$$

Since $\mathcal{F}$ and $\mathcal{G}$ are the union of $\binom{p}{Cs}$ VC classes of dimension Cs, it satisfies $\log N(\varepsilon\|F\|_{2,\mathbb{P}_n},\mathcal{F},\mathbb{P}_n)\lesssim Cs\log(p\vee n)+Cs\log(1/\varepsilon)$.

To estabilsh relation (A.41), note that by $\widehat{g}_{\tau_i}=x_i'\widetilde{\beta}_\tau$ and $\mathrm{E}[f_iv_ix_i]=0$, we have

$$\bar{\mathrm{E}}[f_iv_i\{\tilde{g}_{\tau i}-g_{\tau i}\}]\big|_{\tilde{g}_{\tau i}=\widehat{g}_{\tau i}}=\bar{\mathrm{E}}[f_iv_ig_{\tau i}]=\bar{\mathrm{E}}[f_iv_ir_{g\tau i}]=O(\delta_nn^{-1/2})$$

where the last relation follows from Condition AS(iv). Therefore, since $\widetilde{\beta}_\tau-\beta_\tau\in\mathcal{G}$ with probability $1-o(1)$, using triangle inequality and Lemma 9 together with the entropy bounds for $\mathcal{G}$ we have

$$|\mathbb{E}_n[f_iv_i\{\widehat{g}_{\tau i}-g_{\tau_i}\}]|\leqslant O(\delta_nn^{-1/2})+\sup_{\delta\in\mathcal{G}}|(\mathbb{E}_n-\bar{\mathrm{E}})[f_iv_i\{x_i'\delta+r_{g\tau i}\}]|\lesssim_P\delta_nn^{-1/2}+\sqrt{\frac{s\log(p\vee n)}{n}}\sqrt{\frac{s\log(p\vee n)}{n}}$$

which yields (A.41) under $s^2\log^2(p\vee n)\leqslant\delta_nn$.

To show Condition IQR(iii) relation (A.42) note that

$$\sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[\varphi_\tau(y_i,x_i'\widetilde{\beta}_\tau+d_i\alpha)\widehat{v}_i-\varphi_\tau(y_i,g_{\tau i}+d_i\alpha)v_i\right]\right|$$

$$\leqslant\sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[\varphi_\tau(y_i,x_i'\widetilde{\beta}_\tau+d_i\alpha)\{\widehat{v}_i-v_i\}\right]\right|+\tag{B.44}$$

$$+\sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[\{\varphi_\tau(y_i,x_i'\widetilde{\beta}_\tau+d_i\alpha)-\varphi_\tau(y_i,x_i'\beta_\tau+d_i\alpha)\}v_i\right]\right|+\tag{B.45}$$

$$+\sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[\{\varphi_\tau(y_i,x_i'\beta_\tau+d_i\alpha)-\varphi_\tau(y_i,g_{\tau i}+d_i\alpha)\}v_i\right]\right|.\tag{B.46}$$

To bound (B.44), we write $\widehat{v}_i - v_i = f_i x_i'\{\widetilde{\theta}_\tau - \theta_\tau\} + f_i r_{\theta\tau i}$. Substitute the equation above into (B.44). Again using triangle inequality and Lemma 9 together with the entropy bounds for $\mathcal{F}$ and $\mathcal{G}$ we have

$$\text{(B.44).(i)} = \sup_{\alpha\in\mathcal{A}_\tau} \left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha) f_i x_i'\{\widetilde{\theta}_\tau - \theta_\tau\}\right]\right| \lesssim_P \sqrt{\frac{Cs\log(p\vee n)}{n}}\sqrt{\frac{s\log(p\vee n)}{n}}$$

$$\text{(B.44).(ii)} = \sup_{\alpha\in\mathcal{A}_\tau} \left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha) f_i r_{\theta\tau i}\right]\right| \lesssim_P \sqrt{\frac{Cs\log(p\vee n)}{n}}\sqrt{\frac{s}{n}}$$

To bound (B.45), by Lemma 10, $\|x_i'\{\widetilde{\beta}_\tau - \beta_\tau\}\|_{2,n} \lesssim_P \sqrt{s\log(p\vee n)/n}$, $\bar{\mathrm{E}}[(1\{y_i \leqslant a\} - 1\{y_i \leqslant b\})^2 v_i^2] \leqslant \bar{\mathrm{E}}[\bar{f} v_i^2 |a-b|]$, and $\|v_i^2\|_{2,n} \lesssim_P \{\bar{\mathrm{E}}[v_i^4]\}^{1/2}$, we have

$$\sup_{\alpha\in\mathcal{A}_\tau} \left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\{\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha) - \varphi_\tau(y_i, x_i'\beta_\tau + d_i\alpha)\} v_i\right]\right|$$

$$\lesssim_P \sqrt{\frac{Cs\log(p\vee n)\log n}{n}}\left(\{\bar{\mathrm{E}}[v_i^4]\}^{1/2}\bar{f}\sqrt{\frac{Cs\log(p\vee n)}{n}} + \sqrt{\frac{s\log(p\vee n)}{n}}\{\bar{\mathrm{E}}[v_i^4]\}^{1/2}\right)^{1/2} = o_P(n^{-1/2})$$

provided that $s^3\log^3(p\vee n)\log^2 n \leqslant \delta_n n$. Similarly, to bound (B.46), by Lemma 10 and $\|r_{g\tau i}\|_{2,n} \lesssim_P \sqrt{s/n}$, we have

$$\sup_{\alpha\in\mathcal{A}_\tau} \left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\{\varphi_\tau(y_i, x_i'\beta_\tau + d_i\alpha) - \varphi_\tau(y_i, g_{\tau i} + d_i\alpha)\} v_i\right]\right|$$

$$\lesssim_P \sqrt{\frac{\log n}{n}}\left(\{\bar{\mathrm{E}}[v_i^4]\}^{1/2}\bar{f}\|r_{g\tau i}\|_{2,n}^{1/2} + \{\bar{\mathrm{E}}[v_i^4]\}^{1/2}\sqrt{\frac{\log n}{n}}\right)^{1/2} = o_P(n^{-1/2}).$$

Next we verify the second part of Condition IQR(iii) relation (A.43). To show

$$|\mathbb{E}_n[\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\check{\alpha}_\tau)\widehat{v}_i]| \lesssim_P \delta_n\, n^{-1/2}$$

consider that

$$L_n(\check{\alpha}_\tau) = \frac{\{\mathbb{E}_n[\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\check{\alpha}_\tau)\widehat{v}_i]\}^2}{\mathbb{E}_n[\varphi_\tau^2(y_i, x_i'\widetilde{\beta}_\tau + d_i\check{\alpha}_\tau)\widehat{v}_i^2]} = \min_{\alpha\in\mathcal{A}_\tau} \frac{\{\mathbb{E}_n[\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha)\widehat{v}_i]\}^2}{\mathbb{E}_n[\varphi_\tau^2(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha)\widehat{v}_i^2]}$$

$$\leqslant \frac{1}{\tau^2(1-\tau)^2\mathbb{E}_n[\widehat{v}_i^2]} \min_{\alpha\in\mathcal{A}_\tau} \{\mathbb{E}_n[\varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha)\widehat{v}_i]\}^2$$

Letting $\widehat{\varphi}_i(\alpha) = \varphi_\tau(y_i, x_i'\widetilde{\beta}_\tau + d_i\alpha)$, $\varphi_i(\alpha) = \varphi_\tau(y_i, g_{\tau i} + d_i\alpha)$ we have

$$\begin{aligned}|\mathbb{E}_n[\widehat{\varphi}_i(\alpha)\widehat{v}_i]| &\leqslant |(\mathbb{E}_n - \bar{\mathrm{E}})[\widehat{\varphi}_i(\alpha)\widehat{v}_i - \varphi_i(\alpha)v_i]| + |\bar{\mathrm{E}}[\widehat{\varphi}_i(\alpha)\widehat{v}_i] - \bar{\mathrm{E}}[\varphi_i(\alpha)v_i]| + |\mathbb{E}_n[\varphi_i(\alpha)v_i]| \\ &\lesssim_P \delta_n\, n^{-1/2} + \delta_n|\alpha - \alpha_\tau| + |\mathbb{E}_n[\varphi_i(\alpha)v_i]|\end{aligned}$$

where the bias term $|\bar{\mathrm{E}}[\widehat{\varphi}_i(\alpha)\widehat{v}_i] - \bar{\mathrm{E}}[\varphi_i(\alpha)v_i]| \lesssim_P \delta_n\, n^{-1/2} + \delta_n|\alpha - \alpha_\tau|$ follows from relations (H.74), (H.75), and (H.77) in the Supplementary Appendix. Therefore,

$$\begin{aligned}\frac{\{\ \mathbb{E}_n[\widehat{\varphi}_i(\check{\alpha}_\tau)\widehat{v}_i]\ \}^2}{\mathbb{E}_n[\widehat{v}_i^2]} \leqslant L_n(\check{\alpha}_\tau) &\leqslant \frac{\{\mathbb{E}_n[\widehat{v}_i^2]\}^{-1}}{\tau^2(1-\tau)^2} \min_{\alpha\in\mathcal{A}_\tau}\{\ \mathbb{E}_n[\widehat{\varphi}_i(\alpha)\widehat{v}_i]\ \}^2 \\ &\lesssim_P \frac{\{\mathbb{E}_n[\widehat{v}_i^2]\}^{-1}}{\tau^2(1-\tau)^2} \min_{\alpha\in\{\alpha:|\alpha-\alpha_\tau|\leqslant n^{-1/2}/\delta_n\}}\{\ \delta_n\, n^{-1/2} + \delta_n|\alpha - \alpha_\tau| + |\mathbb{E}_n[\varphi_i(\alpha)v_i]|\ \}^2 \\ &\lesssim_P \frac{\{\mathbb{E}_n[\widehat{v}_i^2]\}^{-1}}{\tau^2(1-\tau)^2}\{\ \delta_n\, n^{-1/2} + \delta_n|\alpha^* - \alpha_\tau| + |\mathbb{E}_n[\varphi_i(\alpha^*)v_i]|\ \}^2\end{aligned}$$

where $\alpha^* \in \arg\min_{\alpha\in\{\alpha:|\alpha-\alpha_\tau|\leqslant n^{-1/2}/\delta_n\}} |\mathbb{E}_n[\varphi_i(\alpha)v_i]|$. It follows that $|\alpha^*-\alpha_\tau| \lesssim_P n^{-1/2}$ and $|\mathbb{E}_n[\varphi_i(\alpha^*)v_i]| \lesssim_P n^{-1}\max_{i\leqslant n}|v_i|$. Therefore, since $\max_{i\leqslant n}|v_i| \lesssim_P n^{1/4}$ by $\bar{\mathrm{E}}[v_i^4] \leqslant C$, we have

$$|\mathbb{E}_n[\widehat{\varphi}_i(\check{\alpha}_\tau)\widehat{v}_i]| \lesssim_P \frac{\delta_n n^{-1/2}}{\tau(1-\tau)}.$$

□

Proof of Theorem 2. The analysis of Step 1 and 2 are identical to the corresponding analysis in the proof of Theorem 1. Define $(\tilde{y}_i;\tilde{x}_i) = (f_i y_i; f_i d_i, f_i x_i)$, since $f_i = f(d_i,x_i)$ and $0 < \underline{f} \leqslant f_i \leqslant \bar{f}$, by Lemma 2 we have $\|x_i'\check{\beta}_\tau - g_{\tau i}\|_{2,n} \lesssim_P \sqrt{s\log(p\vee n)/n}$ and $|\check{\alpha}_\tau - \alpha_\tau| \leqslant \delta_n$. (Note that the verification of the side conditions follows as the verification for Step 1 since $0 < \underline{f} \leqslant f_i \leqslant \bar{f}$.)

Next we construct instruments from the first order conditions of Step 3. Let $\widehat{T}_\tau^*$ denote the variables selected in Steps 1 and 2: $\widehat{T}_\tau^* := \mathrm{support}(\widehat{\beta}_\tau^{(2s)}) \cup \mathrm{support}(\widehat{\theta}_\tau)$. By the first order conditions of the the weighted quantile regression optimization problem, $(\check{\alpha}_\tau, \check{\beta}_\tau)$ are such that there are $s_i \in \partial\rho_\tau(y_i - d_i\check{\alpha}_\tau - x_i'\check{\beta}_\tau)$, $i = 1,\ldots,n$, such that

$$\mathbb{E}_n[s_i f_i(d_i, x_{i\widehat{T}_\tau^*}')'] = 0.$$

Trivially $\mathbb{E}_n[s_i f_i(d_i, x_{i\widehat{T}_\tau^*}')](1,-\widehat{\theta}_\tau')' = 0$ since it is a linear combination of the equations above. Therefore, defining $\widehat{v}_i = f_i(d_i - x_{i\widehat{T}_\tau^*}'\widehat{\theta}_\tau)$, we have $\mathbb{E}_n[s_i\widehat{v}_i] = 0$. Moreover, since $s_i = \varphi_\tau(y_i, d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau)$ if $y_i \neq d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau$,

$$\begin{aligned}
|\mathbb{E}_n[\varphi_\tau(y_i, d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau)\widehat{v}_i]| &\leqslant |\mathbb{E}_n[s_i\widehat{v}_i]| + \mathbb{E}_n[1\{y_i = d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\}|\widehat{v}_i|] \\
&\leqslant \mathbb{E}_n[1\{y_i = d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\}|\widehat{v}_i - v_i|] + \mathbb{E}_n[1\{y_i = d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\}|v_i|] \\
&\leqslant \sqrt{(1+|\widehat{T}_\tau^*|)/n}\|\widehat{v}_i - v_i\|_{2,n} + \max_{i\leqslant n}|v_i|(1+|\widehat{T}_\tau^*|)/n.
\end{aligned}$$

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

$$\widetilde{L}_n(\alpha) = \frac{|\mathbb{E}_n[\varphi_\tau(y_i, d_i\alpha + x_i'\check{\beta}_\tau)\widehat{v}_i]|^2}{\mathbb{E}_n[\{\varphi_\tau(y_i, d_i\alpha + x_i'\check{\beta}_\tau)\}^2\widehat{v}_i^2]},$$

Since $|\widehat{T}_\tau^*| \lesssim_P s$, $\sqrt{s}\|\widehat{v}_i - v_i\|_{2,n} = o_P(1)$, $s^3 \leqslant \delta_n n$, and $\max_{i\leqslant n}|v_i| \lesssim_P n^{1/6}$ by $\bar{\mathrm{E}}[v_i^6] \leqslant C$ we have

$$\sqrt{(1+|\widehat{T}_\tau^*|)/n}\|\widehat{v}_i - v_i\|_{2,n} + \max_{i\leqslant n}|v_i|(1+|\widehat{T}_\tau^*|)/n \lesssim_P \sqrt{s/n}\|\widehat{v}_i - v_i\|_{2,n} + n^{1/6}s/n = o(n^{-1/2}).$$

The result follows by Lemma 5. □

B.1. **Proof of Theorems for Unknown Density.**

Proof of Theorem 3. The proof can be found in the Supplementary Material. □

Proof of Theorem 4. The proof can be found in the Supplementary Material. □

Appendix C. Auxiliary Inequalities

Lemma 6. *Consider $\widehat{\beta}$ and β_0 where $\|\beta_0\|_0 \leqslant s$, and denote $\widehat{\beta}^{(m)}$ as the vector $\widehat{\beta}$ truncated to have only its $m \geqslant s$ largest components. We have that*

$$\begin{aligned} \|\widehat{\beta}^{(m)} - \beta_0\|_1 &\leqslant 2\|\widehat{\beta} - \beta_0\|_1 \\ \|x_i'(\widehat{\beta}^{(2m)} - \beta_0)\|_{2,n} &\leqslant \|x_i'(\widehat{\beta} - \beta_0)\|_{2,n} + \sqrt{\phi_{\max}(m)/m}\|\widehat{\beta} - \beta_0\|_1. \end{aligned}$$

Lemma 7 (Maximal inequality via symmetrization)**.** *Let $Z_1, \ldots, Z_n$ be arbitrary independent stochastic processes and $\mathcal{F}$ a finite set of measurable functions. For any $\tau \in (0, 1/2)$, and $\delta \in (0, 1)$ we have that with probability at least $1 - 4\tau - 4\delta$*

$$\max_{f\in\mathcal{F}} |\mathbb{G}_n(f(Z_i))| \leqslant \left\{ 4\sqrt{2\log(2|\mathcal{F}|/\delta)}\; Q\left(\max_{f\in\mathcal{F}} \sqrt{\mathbb{E}_n[f(Z_i)^2]}, 1-\tau \right) \right\} \vee 2 \max_{f\in\mathcal{F}} Q\left(|\mathbb{G}_n(f(Z_i))|, \frac{1}{2} \right).$$

Lemma 8. *Fix arbitrary vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ with $\max_{i\leqslant n} \|x_i\|_\infty \leqslant K_x$. Let ζ_i $(i = 1, \ldots, n)$ be independent random variables such that $\bar{\mathrm{E}}[|\zeta_i|^q] < \infty$ for some $q \geqslant 4$. Then we have with probability $1 - 8\tau$*

$$\max_{1\leqslant j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2\zeta_i^2]| \leqslant 4\sqrt{\frac{\log(2p/\tau)}{n}} K_x^2 (\bar{\mathrm{E}}[|\zeta_i|^q]/\tau)^{4/q}$$

Let us call a threshold function $x : \mathbb{R}^n \mapsto \mathbb{R}$ k-sub-exchangeable if, for any $v, w \in \mathbb{R}^n$ and any vectors $\tilde{v}, \tilde{w}$ created by the pairwise exchange of the components in v with components in w, we have that $x(\tilde{v}) \vee x(\tilde{w}) \geqslant [x(v) \vee x(w)]/k$. Several functions satisfy this property, in particular $x(v) = \|v\|$ with $k = \sqrt{2}$, $x(v) = \|v\|_\infty$ with $k = 1$, and constant functions with $k = 1$.

Lemma 9 (Exponential inequality for separable empirical process)**.** *Consider a separable empirical process $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n \{f(Z_i) - \mathrm{E}[f(Z_i)]\}$ and the empirical measure $\mathbb{P}_n$ for $Z_1, \ldots, Z_n$, an underlying independent data sequence. Let $K > 1$ and $\tau \in (0, 1)$ be constants, and $e_n(\mathcal{F}, \mathbb{P}_n) = e_n(\mathcal{F}, Z_1, \ldots, Z_n)$ be a k-sub-exchangeable random variable, such that*

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

for some universal constant $c > 1$, then

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

Proof. See [3], Lemma 18 and note that the proof does not use that Z_i's are i.i.d., only independent which was the requirement of Lemma 17 of [3]. The statement then follows by a change of variables of $\epsilon = \tilde{\epsilon}\|F\|_{2,\mathbb{P}_n}$. □

Lemma 10. *Suppose that for all $0 < \varepsilon \leqslant \varepsilon_0$*

$$N(\varepsilon, \mathcal{F}, \mathbb{P}_n) \leqslant (\omega/\varepsilon)^m \text{ and } N(\varepsilon, \mathcal{F}^2, \mathbb{P}_n) \leqslant (\omega/\varepsilon)^m, \tag{C.47}$$

for some ω which can grow with n. Then, as n grows we have

$$\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| \lesssim_P \sqrt{m\log(\omega \vee n)} \left(\sup_{f\in\mathcal{F}} \bar{\mathrm{E}}[f^2] + \sqrt{\frac{m\log(n\vee\omega)}{n}} \left(\sup_{f\in\mathcal{F}} \mathbb{E}_n[f^4] \vee \bar{\mathrm{E}}[f^4] \right)^{1/2} \right)^{1/2}.$$

Proof. The result is derived in [5]. □

Lemma 11. *Let X_i, $i = 1, \ldots, n$, be independent random vectors in $\mathbb{R}^p$ be such that $\sqrt{\mathrm{E}[\max_{1 \leqslant i \leqslant n} \|X_i\|_\infty^2]} \leqslant K$. Let*

$$\delta_n := 2 \left(\bar{C} K \sqrt{k} \log(1+k) \sqrt{\log(p \vee n)} \sqrt{\log n} \right) / \sqrt{n},$$

where $\bar{C}$ is the universal constant. Then,

$$\mathrm{E}\left[\sup_{\|\alpha\|_0 \leqslant k, \|\alpha\|=1} \left| \mathbb{E}_n \left[(\alpha' X_i)^2 - \mathrm{E}[(\alpha' X_i)^2] \right] \right| \right] \leqslant \delta_n^2 + \delta_n \sup_{\|\alpha\|_0 \leqslant k, \|\alpha\|=1} \sqrt{\bar{\mathrm{E}}[(\alpha' X_i)^2]}.$$

Proof. It follows from Theorem 3.6 of [33], see [7] for details. □

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28:253–263, 2008.

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

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

[4] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[5] A. Belloni, V. Chernozhukov, and I. Fernandez-Val. Conditional quantile processes based on series or many regressors. *arXiv:1105.6154*, may 2011.

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

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

[8] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post model selection inference for LAD regression models. *accepted at Biometrika*, 2014.

[9] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[10] A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for logistic regression with a large number of controls. *ArXiv:1304.3969*, 2013.

[11] Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Chris Hansen. Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645*, 2013.

[12] Alexandre Belloni, Victor Chernozhukov, Lie Wang, et al. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

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

[14] X. Chen. Large sample sieve estimatin of semi-nonparametric models. *Handbook of Econometrics*, 6:5559–5632, 2007.

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

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

[17] N. Fenske, T. Kneib, and T. Hothorn. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the Statistical Association*, 106:494–510, 2011.

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

[19] K. Kato. Group Lasso for high dimensional sparse quantile regression models. *arXiv:1103.1458*, 2011.

[20] K. Knight. Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, 26:755–770, 1998.

[21] R. Koenker. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262, 2011.

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

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

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

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

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

[27] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimator? *The Annals of Statistics*, 34(5):2554–2591, 2006.

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

[29] E. L. Lehmann. *Theory of Point Estimation*. New York: Wiley, 1983.

[30] Joseph P. Romano and Azeem M. Shaikh. On the uniform asymptotic validity of subsampling and the bootstrap. *Ann. Statist.*, 40(6):2798–2822, 2012.

[31] Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4):1378–1408, 2007.

[32] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, 59:3434–3447, 2013.

[33] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045, 2008.

[34] R. J. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.

[35] A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.

[36] Sara Anna van de Geer, Peter Bühlmann, and Ya'acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

[37] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, 1996.

[38] Aad W. van der Vaart and Jon A. Wellner. Empirical process indexed by estimated functions. *IMS Lecture Notes-Monograph Series*, 55:234–252, 2007.

[39] Lie Wang. L_1 penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.*, 120:135–151, 2013.

[40] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242, 2014.

[41] S. Zhou. Restricted eigenvalue conditions on subgaussian matrices. *arXiv:0904.4723v2*, 2009.

Supplementary Appendix for "Valid Post-Selection Inference in High-dimensional Approximately Sparse Quantile Regression Models"

C.1. **Proof of Theorems for Unknown Density.**

Proof of Theorem 3. The proof is similar to the proof of Theorem 1 as we will also verify Condition IQR and the result follows by Lemma 5. The requirement on the conditional density function in IQR(i) is assumed in Condition AS. By setting $\iota_{0i} = v_i$ the other moment conditions in IQR(i) are assumed in Condition AS. The analysis of $\widehat{\alpha}_\tau$, $\widetilde{\alpha}_\tau$, $\widehat{\beta}_\tau$ and $\widetilde{\beta}_\tau$ in Step 1 is the same as in Theorem 1. Therefore $\mathcal{A}_\tau$ satisfies the requirement in IQR(ii). Moreover, $|\check{\alpha}_\tau - \alpha_\tau| \lesssim_P \sqrt{s\log(n\vee p)/n}$ satisfies the first part of (A.43), and $\|x_i'\widetilde{\beta}_\tau - g_{\tau i}\|_{2,n} \lesssim_P \sqrt{s\log(n\vee p)/n}$. The second condition in IQR(iv) also follows since

$$\begin{aligned}\|1\{|\epsilon_i| \leqslant |d_i(\alpha_\tau - \check{\alpha}_\tau) + g_{\tau i} - \widehat{g}_i|\}\|_{2,n}^2 &\leqslant \mathbb{E}_n[1\{|\epsilon_i| \leqslant |d_i(\alpha_\tau - \check{\alpha}_\tau)| + |x_i'(\widetilde{\beta}_\tau - \beta_\tau)| + |r_{g\tau i}|\}] \\ &\leqslant \mathbb{E}_n[1\{|\epsilon_i| \leqslant 3|d_i(\alpha_\tau - \check{\alpha}_\tau)|\}] + \mathbb{E}_n[1\{|\epsilon_i| \leqslant 3|x_i'(\widetilde{\beta}_\tau - \beta_\tau)|\}] \\ &\quad + \mathbb{E}_n[1\{|\epsilon_i| \leqslant 3|r_{g\tau i}|\}] \lesssim_P \bar{f}K_x\sqrt{s^2\log(n\vee p)/n}.\end{aligned}$$

Next we establish rates for $\widehat{f}_i$. Under Condition D we have

$$\|f_i - \widehat{f}_i\|_{2,n} \lesssim_P \frac{1}{h}\sqrt{\frac{s\log(n\vee p)}{n}} + h^{\bar{k}} \quad \text{and} \quad \max_{i\leqslant n}|\widehat{f}_i - f_i| \lesssim_P \delta_n \tag{C.48}$$

where $\bar{k}$ depends on the estimator. Let $\mathcal{U}$ denote the set of quantile indices used in the calculation of $\widehat{f}_i$.

Step 2 relies on Post-Lasso with estimated weights. Condition WL(i) and (ii) are implied by Conditions AS. Indeed, the moment conditions in AS imply the first part, and $\Phi^{-1}(1-\gamma/2p) \leqslant \delta_n n^{1/3}$ is implied by $\log(1/\gamma) \lesssim \log(p\vee n)$ and $\log^3 p \leqslant \delta_n n$. The first part of Condition WL(iii) is implied by Lemma 8 under the moment conditions and the growth condition $K_x^4 \log p \leqslant \delta_n n$. Condition WL(iv) follows similarly as in the proof of Theorem 1 using the uniform consistency in (C.48).

The second part of Condition WL(iii) follows from (C.48) and Condition WL(iv) since

$$\max_{j\leqslant p}\mathbb{E}_n[(\widehat{f}_i - f_i)^2 x_{ij}^2 v_i^2] \leqslant \max_{i\leqslant n}|\widehat{f}_i - f_i|^2\left\{\max_{j\leqslant p}(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2 v_i^2] + \max_{j\leqslant p}\bar{\mathrm{E}}[x_{ij}^2 v_i^2]\right\} \lesssim_P \delta_n.$$

The third part of Condition WL(iii) follows from (C.48) and Condition WL(i,ii) since

$$c_r^2 = \mathbb{E}_n[\widehat{f}_i^2 r_{\theta\tau i}^2] \leqslant \max_{i\leqslant n}\{|\widehat{f}_i - f_i| + |f_i|\}\mathbb{E}_n[r_{\theta\tau i}^2] \lesssim_P s/n.$$

To show the fourth part of Condition WL(iii) we note that $\max_{i\leqslant n}\widehat{f}_i^2(f_i + Ch^{\bar{k}})^2 \lesssim_P C$ and $1/\min_{i\leqslant n} f_i^2 \leqslant C$. Letting $\delta_u = \widetilde{\beta}_u - \beta_u$ and $\vartheta_u = \widetilde{\alpha}_u - \alpha_u$, for $u \in \mathcal{U}$, we have

$$\begin{aligned}c_f^2 &= \mathbb{E}_n[(\widehat{f}_i - f_i)^2 v_i^2/f_i^2] \\ &\lesssim_P h^{2\bar{k}}\mathbb{E}_n[v_i^2] + h^{-2}\sum_{u\in\mathcal{U}}\mathbb{E}_n[v_i^2(x_i'\delta_u)^2 + v_i^2 d_i^2\vartheta_u^2 + v_i^2 r_{ui}^2]\end{aligned} \tag{C.49}$$

Conditional on $\{z_i, i = 1, \ldots, n\}$, note the following relations for $u \in \mathcal{U}$

$$\begin{aligned}
\mathbb{E}_n[v_i^2 r_{ui}^2] &\lesssim_P \bar{\mathrm{E}}[v_i^2 r_{ui}^2] = \mathbb{E}_n[r_{ui}^2 \mathrm{E}[v_i^2 \mid z_i]] \leqslant \mathbb{E}_n[r_{ui}^2] \max_{i \leqslant n} \mathrm{E}[v_i^2 \mid z_i] \lesssim s/n \\
\mathbb{E}_n[v_i^2 d_i^2 \vartheta_u^2] &= \mathbb{E}_n[v_i^2 d_i^2]\vartheta_u^2 \leqslant \{\mathbb{E}_n[v_i^4]\mathbb{E}_n[d_i^4]\}^{1/2}\vartheta_u^2 \lesssim_P s \log(p \vee n)/n \\
\mathbb{E}_n[v_i^2 (x_i'\delta_u)^2] &= \mathbb{E}_n[(x_i'\delta_u)^2 \mathrm{E}[v_i^2 \mid z_i]] + (\mathbb{E}_n - \bar{\mathrm{E}})[v_i^2 (x_i'\delta_u)^2] \\
&\leqslant \mathbb{E}_n[(x_i'\delta_u)^2] \max_{i \leqslant n} \mathrm{E}[v_i^2 \mid z_i] + \|\delta_u\|^2 \sup_{\|\delta\|_0 \leqslant \|\delta_u\|_0, \|\delta\|=1} |(\mathbb{E}_n - \bar{\mathrm{E}})[\{v_i x_i'\delta\}^2]|
\end{aligned}$$

To bound the last term we have $\|x_i'\delta_u\|_{2,n}^2 \lesssim_P s \log(n \vee p)/n$ and $\|\delta_u\|_0 \leqslant 2Cs$ with probability $1 - \Delta_n$ by Condition D. Then we apply Lemma 11 with $X_i = v_i x_i$. Thus, we can take $K = \{\mathrm{E}[\max_{i \leqslant n} \|X_i\|_\infty^2]\}^{1/2} \leqslant K_x\{\bar{\mathrm{E}}[\max_{i \leqslant n} v_i^2]\}^{1/2} \lesssim n^{1/8} K_x$ (since $\bar{\mathrm{E}}[v_i^8] \leqslant C$), and $\bar{\mathrm{E}}[(\delta' X_i)^2] \leqslant \mathbb{E}_n[(x_i'\delta)^2] \max_{i \leqslant n} \mathrm{E}[v_i^2 \mid z_i] \leqslant C\phi_{\max}(\|\delta\|_0)\|\delta\|^2$. Therefore,

$$\begin{aligned}
\sup_{\|\delta\|_0 \leqslant 2Cs, \|\delta\|=1} \left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\{v_i x_i'\delta\}^2\right]\right| &\lesssim_P \left\{ \tfrac{K_x^2 n^{1/4} s \log^3 n \log(p \vee n)}{n} + \sqrt{\tfrac{K_x^2 n^{1/4} s \log^3 n \log(p \vee n)}{n}} \phi_{\max}(2Cs) \right\} \\
&\lesssim \left\{ \tfrac{K_x \log^3 n}{n^{1/4}} \tfrac{K_x s \log(p \vee n)}{n^{1/2}} + \sqrt{\tfrac{K_x \log^3 n}{n^{1/4}} \tfrac{K_x s \log(p \vee n)}{n^{1/2}}} \phi_{\max}(2Cs) \right\}
\end{aligned}$$

under the conditions $K_x^4 \leqslant \delta_n n^{4/q}$, $q > 4$, and $K_x^2 s^2 \log^2(p \vee n) \leqslant \delta_n n$, and $\phi_{\max}(s/\delta_n)$ being bounded from above with probability $1 - \Delta_n$ by Condition SE. Therefore,

$$c_f^2 \lesssim_P \frac{s \log(n \vee p)}{h^2 n} + h^{2\bar{k}}.$$

Under Condition WL, by Lemma 4 we have

$$\begin{aligned}
\|\widetilde{\theta}_\tau\|_0 &\lesssim_P \frac{n^2\{c_f^2 + c_r^2\}}{\lambda^2} + s \lesssim \widetilde{s}_{\theta\tau} := s + \frac{ns \log(n \vee p)}{h^2\lambda^2} + \left(\frac{nh^{\bar{k}}}{\lambda}\right)^2 \text{ and} \\
\|x_i'(\widetilde{\theta}_\tau - \theta_\tau)\|_{2,n} &\lesssim_p \frac{1}{h}\sqrt{\frac{s \log(n \vee p)}{n}} + h^{\bar{k}} + \frac{\lambda\sqrt{s}}{n}
\end{aligned}$$

where we used that $\phi_{\max}(\widetilde{s}_{\theta\tau}/\delta_n) \leqslant C$, and that $\lambda \geqslant \sqrt{n}\Phi^{-1}(1 - \gamma/2p) \sim \sqrt{n \log(p/\gamma)}$ so that

$$\sqrt{\frac{\widetilde{s}_{\theta\tau} \log p}{n}} \lesssim \frac{1}{h}\sqrt{\frac{s \log p}{n}} + h^{\bar{k}}.$$

For convenience we write $\tilde{x}_i = (d_i, x_i')'$ and we will consider the following classes of functions

$$\begin{aligned}
\mathcal{K} &= \{x_i'\beta : \|\beta\|_0 \leqslant Cs,\ \|\beta - \beta_\tau\| \leqslant C\sqrt{s \log(p \vee n)/n}\} \\
\mathcal{F} &= \{\tau - 1\{y_i \leqslant x_i'\beta + d_i\alpha\} : \|\beta\|_0 \leqslant Cs,\ \|\beta - \beta_\tau\| \leqslant C\sqrt{s \log(p \vee n)/n}, |\alpha - \alpha_\tau| \leqslant \delta_n\} \\
\mathcal{G} &= \{x_i'\delta : \|x_i'\delta\|_{2,n} \leqslant C\{\tfrac{1}{h}\sqrt{s \log p/n} + h^{\bar{k}} + \tfrac{\lambda\sqrt{s}}{n}\}, \|\delta\|_0 \leqslant C\widetilde{s}_{\theta\tau}\} \\
\mathcal{J} &= \left\{ \tilde{f}_i : \begin{array}{l} \|\tilde{\eta}_u\|_0 \leqslant Cs, \|\tilde{x}_i'\tilde{\eta}_u - Q(u \mid d_i, z_i)\|_{2,n} \leqslant C\sqrt{s \log(p \vee n)/n}, \\ \|\tilde{x}_i'\tilde{\eta}_u - Q(u \mid d_i, z_i)\|_\infty \leqslant \delta_n h, u \in \mathcal{U} \end{array} \right\}
\end{aligned} \tag{C.50}$$

We have that $\mathcal{K}$ and $\mathcal{F}$ are the union of $\binom{p}{Cs}$ VC classes of dimension Cs and $\mathcal{G}$ is the union of $\binom{p}{\widetilde{s}_{\theta\tau}}$ VC classes of dimension $C\widetilde{s}_{\theta\tau}$. Thus, we have that $\log N(\varepsilon\|F\|_{2,\mathbb{P}_n}, \mathcal{F}, \mathbb{P}_n) \lesssim Cs \log p + Cs \log(1/\varepsilon)$ and $\log N(\varepsilon\|G\|_{2,\mathbb{P}_n}, \mathcal{G}, \mathbb{P}_n) \lesssim C\widetilde{s}_{\theta\tau} \log p + C\widetilde{s}_{\theta\tau} \log(1/\varepsilon)$ where $\|F\|_{2,\mathbb{P}_n} \leqslant 1$ and $G(y, d, x) = \max_{\delta \in \mathcal{G}} |x_i'\delta|$. Under the choice of bandwidth h in Condition D, we have $CK_x\sqrt{s^2 \log(n \vee p)/n} \leqslant \delta_n h$, and the functions in $\mathcal{J}$ are uniformly bounded above and below. Moreover, $\mathcal{J}$ is the union of $\binom{p}{Cs}^{\bar{k}}$ VC classes of dimension $C's$ so that $\log N(\varepsilon\|J\|_{2,\mathbb{P}_n}, \mathcal{J}, \mathbb{P}_n) \lesssim Cs \log p + Cs \log(1/\varepsilon)$ where $J(y, d, z) = \sup_{\tilde{f} \in \mathcal{J}} |\tilde{f}(y, d, z)|$.

Next we provide bounds required by IQR(iii). We have

$$\begin{aligned}\{\mathbb{E}_n[(\widehat{v}_i - v_i)^2]\}^{1/2} &\leqslant \{\mathbb{E}_n[\{(\widehat{f}_i - f_i)(d_i - x_i'\widetilde{\theta}_\tau)\}^2]\}^{\frac{1}{2}} + \{\mathbb{E}_n[\{f_i x_i'(\theta_\tau - \widetilde{\theta}_\tau)\}^2]\}^{\frac{1}{2}} + \{\mathbb{E}_n[\{f_i r_{\theta\tau i}\}^2]\}^{\frac{1}{2}} \\ &\lesssim_P \{\mathbb{E}_n[\{(\widehat{f}_i - f_i)v_i/f_i\}^2]\}^{\frac{1}{2}} + \{\mathbb{E}_n[\{(\widehat{f}_i - f_i)x_i'(\theta_\tau - \widetilde{\theta}_\tau)\}^2]\}^{\frac{1}{2}} \\ &+ \{\mathbb{E}_n[\{(\widehat{f}_i - f_i)r_{\theta\tau i}\}^2]\}^{\frac{1}{2}} + \max_{i\leqslant n} f_i\{\|x_i'(\theta_\tau - \widetilde{\theta}_\tau)\|_{2,n} + \|r_{\theta\tau i}\|_{2,n}\} \\ &\lesssim_P c_f + \max_{i\leqslant n}|\widehat{f}_i - f_i|\ \|x_i'(\theta_\tau - \widetilde{\theta}_\tau)\|_{2,n} \\ &+ \tfrac{1}{h}\sqrt{\tfrac{s\log(n\vee p)}{n}} + h^{\bar{k}} + \tfrac{\lambda}{n}\sqrt{s}.\end{aligned}$$

Therefore, since $\max_{i\leqslant n}|\widehat{v}_i - v_i| \lesssim_P \delta_n$ and $\max_{i\leqslant n}|v_i| \lesssim_P n^{1/6}$ since $\bar{\mathrm{E}}[v_i^6] \leqslant C$, we have

$$\begin{aligned}\max_{i\leqslant n}\{1 + |v_i| + |\widehat{v}_i - v_i|\}^{1/2}\|g_{\tau i} - x_i'\widetilde{\beta}_\tau\|_{2,n} &\lesssim_P n^{-1/4}\left\{\tfrac{\max_{i\leqslant n}\{1+|v_i|+|\widehat{v}_i - v_i|\}}{n^{1/6}}\tfrac{s\log(p\vee n)}{n^{1/3}}\right\}^{1/2} \\ \{\mathbb{E}_n[(\widehat{v}_i - v_i)^2]\}^{1/2} &\lesssim_P (1/h)\{s\log(p\vee n)/n\}^{1/2} + h^{\bar{k}} + \tfrac{\lambda}{n}\sqrt{s} \lesssim \delta_n, \\ \{\mathbb{E}_n[(\widehat{v}_i - v_i)^2]\}^{1/2}\|g_{\tau i} - x_i'\widetilde{\beta}_\tau\|_{2,n} &\lesssim_P \left\{\tfrac{1}{h}\sqrt{\tfrac{s\log(p\vee n)}{n}} + h^{\bar{k}}\right\}\sqrt{\tfrac{s\log(p\vee n)}{n}} \\ &\lesssim n^{-1/2}\left\{\tfrac{1}{h}\tfrac{s\log(p\vee n)}{n^{1/2}} + h^{\bar{k}}\sqrt{s\log(p\vee n)}\right\}\end{aligned}$$

The last condition in (A.40) follows from Lemma 7 and the entropy bounds on $\mathcal{K}$

$$\begin{aligned}|\mathbb{E}_n[f_i v_i\{x_i'\widetilde{\beta}_\tau - g_{\tau i}\}| &\lesssim_P \sup_{w\in\mathcal{K}}|\mathbb{E}_n[f_i\iota_{0i}\{w_i - g_{\tau i}\}]| \lesssim_P \sqrt{\tfrac{s\log(n\vee p)}{n}}\sqrt{\tfrac{\{\max_{i\leqslant n} v_i^2 + \mathrm{E}[v_i^2]\}s\log(p\vee n)}{n}} \\ &\lesssim_P n^{-1/2}\left\{\tfrac{\max_{i\leqslant n} v_i^2 + \mathrm{E}[v_i^2]}{n^{1/3}}\tfrac{s^2\log^2(n\vee p)}{n^{2/3}}\right\}^{1/2}\end{aligned}$$

Next we verify (A.42). Let $\widehat{\varphi}_i(\alpha) = \varphi_\tau(y_i, x_i'\widehat{\beta}_\tau + d_i\alpha)$, $\varphi_i(\alpha) = \varphi_\tau(y_i, x_i'\beta_\tau + d_i\alpha)$. To show Condition IQR(ii) note that

$$\begin{aligned}&\sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\widehat{\varphi}_i(\alpha)\widehat{v}_i - \varphi_\tau(y_i, g_{\tau i} + d_i\alpha)v_i\right]\right| \\ &\quad\leqslant \sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\widehat{\varphi}_i(\alpha)(\widehat{v}_i - v_i)\right]\right| + \qquad (C.51) \\ &\quad+ \sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\{\widehat{\varphi}_i(\alpha) - \varphi_i(\alpha)\}v_i\right]\right| + \qquad (C.52) \\ &\quad+ \sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\{\varphi_i(\alpha) - \varphi_\tau(y_i, g_{\tau i} + d_i\alpha)\}v_i\right]\right|. \qquad (C.53)\end{aligned}$$

To bound (C.51), we write $\widehat{v}_i - v_i = \widehat{v}_i - \frac{f_i}{\widehat{f}_i}\widehat{v}_i + \frac{f_i}{\widehat{f}_i}\widehat{v}_i - v_i = \widehat{v}_i(\widehat{f}_i - f_i)/\widehat{f}_i + f_i x_i'\{\widetilde{\theta}_\tau - \theta_\tau\} + f_i r_{\theta\tau i}$. Substitute the equation above into (C.51) and using the triangle inequality we have

$$\begin{aligned}(C.51) \quad &\leqslant \sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\widehat{\varphi}_i(\alpha)(d_i - x_i'\widetilde{\theta}_\tau)(\widehat{f}_i - f_i)\right]\right| + \sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\widehat{\varphi}_i(\alpha)f_i x_i'\{\widetilde{\theta}_\tau - \theta_\tau\}\right]\right| \\ &\quad + \sup_{\alpha\in\mathcal{A}_\tau}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[\widehat{\varphi}_i(\alpha)f_i r_{\theta\tau i}\right]\right|\end{aligned}$$

Recall that $f_{\epsilon|d,z}(0 \mid d, z) = f(d, z)$ and $r_{\theta\tau} = r_{\theta\tau}(d, z) = m(z) - x'\theta_\tau$. We consider the following combinations of $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{J}$:

$$\begin{aligned}\mathcal{H}^1 &= \{(y, d, z) \mapsto w_1(y, d, z)\{d - x'\theta_\tau - w_2(y, d, z)\}\{w_3(y, d, x) - f(d, z)\} : w_1 \in \mathcal{F}, w_2 \in \mathcal{G}, w_3 \in \mathcal{J}\} \\ \mathcal{H}^2 &= \{(y, d, z) \mapsto w_1(y, d, z)f(d, z)w_2(y, d, z)\ : w_1 \in \mathcal{F}, w_2 \in \mathcal{G}\} \\ \mathcal{H}^3 &= \{(y, d, z) \mapsto w_1(y, d, z)f(d, z)r_{\theta\tau}(d, z)\ : w_1 \in \mathcal{F}\}\end{aligned}$$

Consider the event $\Omega := \{\ \widehat{f}_i \in \mathcal{J},\ x_i'\widetilde{\theta}_\tau \in \mathcal{G},\ \tau - 1\{y_i \leqslant d_i\alpha + x_i'\widetilde{\beta}_\tau\} \in \mathcal{F}$ for all $\alpha \in \mathcal{A}_\tau\ \}$. Under Ω we have

$$(C.51) \leqslant \sup_{w\in\mathcal{H}^1}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[w(y_i, d_i, z_i)\right]\right| + \sup_{w\in\mathcal{H}^2}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[w(y_i, d_i, z_i)\right]\right| + \sup_{w\in\mathcal{H}^3}\left|(\mathbb{E}_n - \bar{\mathrm{E}})\left[w(y_i, d_i, z_i)\right]\right|$$

By Lemma 9 together with entropy bounds based on the entropy bounds for $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{J}$, we have

$$\begin{aligned}
\sup_{w\in\mathcal{H}^1}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[w(y_i,d_i,z_i)\right]\right| &\lesssim_P \sqrt{\tfrac{\widetilde{s}_{\theta\tau}\log(p\vee n)}{n}}\sup_{w\in\mathcal{H}^1}\{(\mathbb{E}_n\vee\bar{\mathrm{E}})[w^2]\}^{1/2}\\
&\lesssim \sqrt{\tfrac{\widetilde{s}_{\theta\tau}\log(p\vee n)}{n}}\sup_{w\in\mathcal{H}^1}\{(\mathbb{E}_n\vee\bar{\mathrm{E}})[\{v_i^2+r_{\theta\tau i}^2+(x_i'\delta)^2\}(\tilde{f}_i-f_i)^2]\}^{1/2}\\
&\lesssim n^{-1/2}\left\{\tfrac{1}{h}\tfrac{\sqrt{\widetilde{s}_{\theta\tau}s\log(n\vee p)}}{\sqrt{n}}+h^{\bar{k}}\sqrt{\widetilde{s}_{\theta\tau}\log(n\vee p)}+\tfrac{\lambda}{n}\sqrt{\widetilde{s}_{\theta\tau}s\log(n\vee p)}\right\}\\
\sup_{w\in\mathcal{H}^2}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[w(y_i,d_i,z_i)\right]\right| &\lesssim_P \sqrt{\tfrac{\widetilde{s}_{\theta\tau}\log(p\vee n)}{n}}\bar{f}\sup_{\delta\in\mathcal{G}}\{\mathbb{E}_n[(x_i'\delta)^2]\}^{1/2}\\
&\lesssim n^{-1/2}\left\{\tfrac{1}{h}\tfrac{\sqrt{\widetilde{s}_{\theta\tau}s\log p}}{\sqrt{n}}+h^{\bar{k}}\sqrt{\widetilde{s}_{\theta\tau}\log(n\vee p)}+\tfrac{\lambda}{n}\sqrt{\widetilde{s}_{\theta\tau}s\log(n\vee p)}\right\}\\
\sup_{w\in\mathcal{H}^3}\left|(\mathbb{E}_n-\bar{\mathrm{E}})\left[w(y_i,d_i,z_i)\right]\right| &\lesssim_P \sqrt{\tfrac{s\log(p\vee n)}{n}}\bar{f}\{\mathbb{E}_n[r_{\theta\tau i}^2]\}^{1/2}\lesssim n^{-1/2}\left\{\tfrac{Cs^2\log(p\vee n)}{n}\right\}^{1/2}
\end{aligned}$$

where we used that $|w_1|\leqslant 1$ for $w_1\in\mathcal{F}$, $\tilde{f}_i$ and f_i are uniformly bounded and (C.49). Plugging in the definition of $\widetilde{s}_{\theta\tau}$ we require the following conditions to hold:

$$\begin{array}{lll}
h^{\bar{k}}\sqrt{s\log(n\vee p)}\leqslant\delta_n, & h^{\bar{k}-1}\sqrt{s\log(n\vee p)}\frac{\sqrt{n\log(n\vee p)}}{\lambda}\leqslant\delta_n, & h^{2\bar{k}}\sqrt{n}\frac{\sqrt{n\log(n\vee p)}}{\lambda}\leqslant\delta_n\\
\frac{s^2\log^2(n\vee p)}{nh^2}\leqslant\delta_n, & \frac{s^2\log^3(n\vee p)}{h^4\lambda^2}\leqslant\delta_n, & \frac{\lambda}{n}s\sqrt{\log(n\vee p)}\leqslant\delta_n.
\end{array}$$

The bounds of (C.52) and (C.53) follows as in the proof of Theorem 1 (since these are not impacted by the estimation of density function). The verification of Condition IQR(iii),

$$|\mathbb{E}_n[\varphi_\tau(y_i,x_i'\widetilde{\beta}_\tau+d_i\check{\alpha}_\tau)\widehat{v}_i]|\leqslant\delta_n\ n^{-1/2},$$

also follows as in the proof of Theorem 1.

The consistency of $\widehat{\sigma}_{1n}$ follows from $\|\tilde{v}_i-v_i\|_{2,n}\to_P 0$ and the moment conditions. The consistency of $\widehat{\sigma}_{3,n}$ follow from Lemma 5. Next we show the consistency of $\widehat{\sigma}_{2n}^2=\{\mathbb{E}_n[\check{f}_i^2(d_i,x_{i\check{T}}')'(d_i,x_{i\check{T}}')]\}_{11}^{-1}$. Because $f_i\geqslant\underline{f}$, sparse eigenvalues of size $\ell_n s$ are bounded away from zero and from above with probability $1-\Delta_n$, and $\max_{i\leqslant n}|\widehat{f}_i-f_i|=o_P(1)$ by Condition D, we have

$$\{\mathbb{E}_n[\widehat{f}_i^2(d_i,x_{i\check{T}}')'(d_i,x_{i\check{T}}')]\}_{11}^{-1}=\{\mathbb{E}_n[f_i^2(d_i,x_{i\check{T}}')'(d_i,x_{i\check{T}}')]\}_{11}^{-1}+o_P(1).$$

So that $\widehat{\sigma}_{2n}-\tilde{\sigma}_{2n}\to_P 0$ for

$$\tilde{\sigma}_{2n}^2=\{\mathbb{E}_n[f_i^2(d_i,x_{i\check{T}}')'(d_i,x_{i\check{T}}')]\}_{11}^{-1}=\{\mathbb{E}_n[f_i^2d_i^2]-\mathbb{E}_n[f_i^2d_ix_{i\check{T}}'][\mathbb{E}_n[f_i^2x_{i\check{T}}x_{i\check{T}}']\}^{-1}\mathbb{E}_n[f_i^2x_{i\check{T}}d_i]\}^{-1}.$$

Next define $\check{\theta}_\tau[\check{T}]=\{\mathbb{E}_n[f_i^2x_{i\check{T}}x_{i\check{T}}']\}^{-1}\mathbb{E}_n[f_i^2x_{i\check{T}}d_i]$ which is the least squares estimator of regressing f_id_i on $f_ix_{i\check{T}}$. Let $\check{\theta}_\tau$ denote the associated p-dimensional vector. By definition $f_ix_i'\theta_\tau=f_id_i-f_ir_{\theta\tau}-v_i$, so that

$$\begin{aligned}
\tilde{\sigma}_{2n}^{-2} &=\mathbb{E}_n[f_i^2d_i^2]-\mathbb{E}_n[f_i^2d_ix_i'\check{\theta}_\tau]\\
&=\mathbb{E}_n[f_i^2d_i^2]-\mathbb{E}_n[f_id_if_ix_i'\theta_\tau]-\mathbb{E}_n[f_id_if_ix_i'(\check{\theta}_\tau-\theta_\tau)]\\
&=\mathbb{E}_n[f_id_iv_i]-\mathbb{E}_n[f_id_if_ir_{\theta\tau i}]-\mathbb{E}_n[f_id_if_ix_i'(\check{\theta}_\tau-\theta_\tau)]\\
&=\mathbb{E}_n[v_i^2]+\mathbb{E}_n[v_if_im_\tau(z_i)]-\mathbb{E}_n[f_id_if_ir_{\theta\tau i}]-\mathbb{E}_n[f_id_if_ix_i'(\check{\theta}-\theta_0)]
\end{aligned}$$

We have that $|\mathbb{E}_n[f_iv_ix_i'\theta_\tau]|=o_P(\delta_n)$ since $\bar{\mathrm{E}}[(v_if_im_\tau(z_i))^2]\leqslant\bar{\mathrm{E}}[v_i^2f_i^2d_i^2]\leqslant\bar{f}^2\{\bar{\mathrm{E}}[v_i^4]\bar{\mathrm{E}}[d_i^4]\}^{1/2}\leqslant C$ and $\bar{\mathrm{E}}[f_iv_ix_i'\theta_\tau]=0$. Moreover, $\mathbb{E}_n[f_id_if_ir_{\theta\tau i}]\leqslant\bar{f}_i^2\|d_i\|_{2,n}\|r_{\theta\tau i}\|_{2,n}=o_P(\delta_n)$, $|\mathbb{E}_n[f_id_if_ix_i'(\check{\theta}-\theta_\tau)]|\leqslant\|d_i\|_{2,n}\|f_ix_i'(\check{\theta}_\tau-\theta_\tau)\|_{2,n}=o_P(\delta_n)$ since $|\check{T}|\lesssim_P\widehat{s}_m+s$ and $\mathrm{support}(\widehat{\theta}_\tau)\subset\check{T}$.

□

Proof of Theorem 4. The analysis of Step 1 and 2 are identical to the corresponding analysis for Algorithm 1'. Let $\widehat{T}_\tau^*$ denote the variables selected in Step 1 and 2: $\widehat{T}_\tau^* = \text{support}(\widehat{\beta}_\tau^{(2s)}) \cup \text{support}(\widehat{\theta}_\tau)$. Using the same arguments as in the proof of Theorem 3, we have

$$|\widehat{T}_\tau^*| \lesssim_P \widehat{s}_\tau^* = s + \frac{ns\log p}{h^2\lambda^2} + \left(\frac{nh^{\bar{k}}}{\lambda}\right)^2.$$

Next we establish preliminary rates for $\check{\beta}_\tau$ and $\check{\alpha}_\tau$. Note that since $\widehat{f}_i$ is a positive function of (d_i, z_i), all the results in Section A.1 apply for $(\tilde{y}_i, \tilde{x}_i) = (\widehat{f}_i y_i, \widehat{f}_i(d_i, x_i')')$ since these results are conditional on (d_i, z_i')'. For $\eta_\tau = (\alpha_\tau, \beta_\tau')'$, $\widehat{\eta}_\tau = (\widehat{\alpha}_\tau, \widehat{\beta}_\tau^{(2s)\prime})'$ and $\check{\eta}_\tau = (\check{\alpha}_\tau, \check{\beta}_\tau')'$ be the solution of

$$\check{\eta}_\tau \in \arg\min_\eta \mathbb{E}_n[\widehat{f}_i \rho_\tau(y_i - (d_i, x'_{i\widehat{T}_\tau^*})\eta)]$$

where $\widehat{f}_i = \widehat{f}_i(d_i, x_i) \geqslant 0$. By definition $\text{support}(\widehat{\beta}_\tau) \subset \widehat{T}_\tau^*$ so that

$$\mathbb{E}_n[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\check{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}] \leqslant \mathbb{E}_n[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\widehat{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}]$$

Therefore we have

$$\begin{aligned}\bar{\mathrm{E}}[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\check{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}] &\leqslant |(\mathbb{E}_n - \bar{\mathrm{E}})[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\check{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}]| \\ &\quad + \mathbb{E}_n[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\widehat{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}]\end{aligned} \tag{C.54}$$

To bound the first term in (C.54) consider the class of functions

$$\mathcal{H} = \{\rho_\tau(y_i - (d_i, x_i')\eta) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau) : \|\eta\|_0 \leqslant C'\widehat{s}_\tau^*, \|(d_i, x_i')(\eta - \eta_\tau)\|_{2,n} \leqslant C\sqrt{\widehat{s}_\tau^* \log p/n}\}$$

Note that $\widehat{f}_i$ is constructed based on the class of functions $\mathcal{J}$ defined in (C.50) which is the union of $\binom{p}{Cs}^2$ uniformly bounded VC classes of dimension $C's$. Therefore,

$$\sup_{\eta\in\mathcal{H}} |(\mathbb{E}_n - \bar{\mathrm{E}})[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\eta) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}]| \lesssim_P \sqrt{\frac{\widehat{s}_\tau^* \log(n \vee p)}{n}}\sqrt{\frac{\widehat{s}_\tau^* \log p}{n}}.$$

To bound the last term in (C.54) let $\delta = \widehat{\eta}_\tau - \eta_\tau$, and note that, conditional on (d_i, x_i'), since $\|(d_i, x_i')'\delta\|_{2,n} \lesssim_P \sqrt{s\log(p \vee n)/n}$, $\|r_{g\tau i}\|_{2,n} \lesssim_P \sqrt{s/n}$ and $\max_{i\leqslant n} \widehat{f}_i \wedge \widehat{f}_i^{-1} \lesssim_P 1$, by Lemma 15 we have

$$\mathbb{E}_n[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\widehat{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}] \lesssim_P \frac{s\log(p \vee n)}{n}.$$

Similarly, Lemma 13 with $(\tilde{y}_i; \tilde{x}_i) := (\widehat{f}_i y_i; \widehat{f}_i d_i, \widehat{f}_i x_i)$, implies that for $\delta = \check{\eta}_\tau - \eta_\tau$,

$$\|(d_i, x_i')\delta\|_{2,n}^2 \wedge \{\bar{q}_A\|(d_i, x_i')\delta\|_{2,n}\} \lesssim \bar{\mathrm{E}}[\widehat{f}_i\{\rho_\tau(y_i - (d_i, x_i')\check{\eta}_\tau) - \rho_\tau(y_i - (d_i, x_i')\eta_\tau)\}] + \sqrt{\frac{\widehat{s}_\tau^*}{n}}\frac{\|(d_i, x_i')\delta\|_{2,n}}{\sqrt{\phi_{\min}(\widehat{s}_\tau^*)}}.$$

Combining these relations with $1/\phi_{\min}(\widehat{s}_\tau^*) \lesssim_P 1$ by Condition D, we have

$$\|(d_i, x_i')'\delta\|_{2,n}^2 \wedge \{\bar{q}_A\|(d_i, x_i')\delta\|_{2,n}\} \lesssim_P \sqrt{\frac{\widehat{s}_\tau^*}{n}}\|(d_i, x_i')\delta\|_{2,n} + \frac{\widehat{s}_\tau^* \log p}{n}$$

which leads to $\|(d_i, x_i')(\check{\eta}_\tau - \eta_\tau)\|_{2,n} \lesssim_P \sqrt{\frac{\widehat{s}_\tau^* \log p}{n}}$.

Next we construct instruments from the first order conditions of Step 3. By the first order conditions for $(\check{\alpha}_\tau, \check{\beta}_\tau)$ in the weighted quantile regression we have for $s_i \in \partial\rho_\tau(y_i - d_i\check{\alpha}_\tau - x_i'\check{\beta}_\tau)$ that

$$\mathbb{E}_n[s_i \widehat{f}_i(d_i, x'_{i\widehat{T}_\tau^*})'] = 0.$$

Since $s_i = \varphi_\tau(y_i, d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau)$ if $y_i \neq d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau$, by taking linear combination of the equation above $(1, -\widetilde{\theta}_\tau)$ and defining $\widehat{v}_i = \widehat{f}_i(d_i - x'_{i\widehat{T}^*_\tau}\widetilde{\theta}_\tau)$ we have

$$\begin{aligned}|\mathbb{E}_n[\varphi_\tau(y_i, d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau)\widehat{v}_i]| &\leqslant |\mathbb{E}_n[s_i\widehat{v}_i]| + \mathbb{E}_n[1\{y_i = d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\}|\widehat{v}_i|] \\ &\leqslant \mathbb{E}_n[1\{y_i = d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\}|\widehat{v}_i - v_i|] + \mathbb{E}_n[1\{y_i = d_i\check{\alpha}_\tau + x_i'\check{\beta}_\tau\}|v_i|] \\ &\leqslant \sqrt{(1 + |\widehat{T}^*_\tau|)/n}\|\widehat{v}_i - v_i\|_{2,n} + \max_{i\leqslant n}|v_i|(1 + |\widehat{T}^*_\tau|)/n.\end{aligned}$$

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

$$\widetilde{L}_n(\alpha) = \frac{|\mathbb{E}_n[\varphi_\tau(y_i, d_i\alpha + x_i'\check{\beta}_\tau)\widehat{v}_i]|^2}{\mathbb{E}_n[\{\varphi_\tau(y_i, d_i\alpha + x_i'\check{\beta}_\tau)\}^2\widehat{v}_i^2]},$$

and we have $\widetilde{L}_n(\check{\alpha}_\tau) = o_P(n^{-1/2})$ since $|\widehat{T}^*_\tau| \lesssim_P \widehat{s}^*_\tau$, provided that $\sqrt{\widehat{s}^*_\tau}\|\widehat{v}_i - v_i\|_{2,n} = o_P(1)$, and $\max_{i\leqslant n}|v_i| \lesssim_P n^{1/4}$ by $\bar{\mathrm{E}}[v_i^4] \leqslant C$.

The remaining growth conditions required to apply Lemma 5 follow from the same requirements used in the proof of Theorem 3

$$\begin{array}{lll} h^{\bar{k}}\sqrt{s\log(n\vee p)} \leqslant \delta_n, & h^{\bar{k}-1}\sqrt{s\log(n\vee p)}\frac{\sqrt{n\log(n\vee p)}}{\lambda} \leqslant \delta_n, & h^{2\bar{k}}\sqrt{n}\frac{\sqrt{n\log(n\vee p)}}{\lambda} \leqslant \delta_n \\ \frac{s^2\log^2(n\vee p)}{nh^2} \leqslant \delta_n, & \frac{s^2\log^3(n\vee p)}{h^4\lambda^2} \leqslant \delta_n, & \frac{\lambda}{n}s\sqrt{\log(n\vee p)} \leqslant \delta_n. \end{array}$$

(Note that the additional condition required by the analysis

$$\frac{\widehat{s}^*_\tau\log(n\vee p)}{\sqrt{n}} \lesssim_P \frac{s\log(n\vee p)}{\sqrt{n}} + \frac{s\log^{3/2}(n\vee p)}{h^2\lambda}\frac{\sqrt{n\log(n\vee p)}}{\lambda} + h^{2\bar{k}}\sqrt{n}\frac{n\log(n\vee p)}{\lambda^2} \leqslant \delta_n$$

is implied by the previous requirements.)

The consistent estimation of σ_n follows as in the proof of Theorem 3. □

Appendix D. Auxiliary Inequalities

Proof of Lemma 6. The first inequality follows from the triangle inequality

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

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

By the triangle inequality we have

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

Note that for integer $k \geqslant 2$, $\|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_0 \leqslant m$ and $\widehat{\beta} - \widehat{\beta}^{(2m)} = \sum_{k\geqslant 3}\{\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\}$. Moreover, given the monotonicity of the components, $\|\widehat{\beta}^{(km+m)} - \widehat{\beta}^{(km)}\| \leqslant \|\widehat{\beta}^{(km)} - \widehat{\beta}^{(km-m)}\|_1/\sqrt{m}$.

Then, we have

$$\begin{aligned}\|x_i'(\widehat{\beta}-\widehat{\beta}^{(2m)})\|_{2,n} &= \|x_i'\textstyle\sum_{k\geqslant 3}\{\widehat{\beta}^{(km)}-\widehat{\beta}^{(km-m)}\}\|_{2,n}\\ &\leqslant \textstyle\sum_{k\geqslant 3}\|x_i'\{\widehat{\beta}^{(km)}-\widehat{\beta}^{(km-m)}\}\|_{2,n}\\ &\leqslant \sqrt{\phi_{\max}(m)}\textstyle\sum_{k\geqslant 3}\|\widehat{\beta}^{(km)}-\widehat{\beta}^{(km-m)}\|\\ &\leqslant \sqrt{\phi_{\max}(m)}\textstyle\sum_{k\geqslant 2}\frac{\|\widehat{\beta}^{(km)}-\widehat{\beta}^{(km-m)}\|_1}{\sqrt{m}}\\ &= \sqrt{\phi_{\max}(m)}\frac{\|\widehat{\beta}-\widehat{\beta}^{(m)}\|_1}{\sqrt{m}}\\ &\leqslant \sqrt{\phi_{\max}(m)}\frac{\|\widehat{\beta}-\beta_0\|_1}{\sqrt{m}}.\end{aligned}$$

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

Lemma 12 (Moderate Deviation Inequality for Maximum of a Vector)**.** *Suppose that*

$$\mathcal{S}_j=\frac{\sum_{i=1}^n U_{ij}}{\sqrt{\sum_{i=1}^n U_{ij}^2}},$$

where U_{ij} are independent variables across i with mean zero. We have that

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathcal{S}_j|>\Phi^{-1}(1-\gamma/2p)\right)\leqslant\gamma\left(1+\frac{A}{\ell_n^3}\right),$$

where A is an absolute constant, provided that for $\ell_n>0$

$$0\leqslant\Phi^{-1}(1-\gamma/(2p))\leqslant\frac{n^{1/6}}{\ell_n}\min_{1\leqslant j\leqslant p}M[U_j]-1,\quad M[U_j]:=\frac{\left(\frac{1}{n}\sum_{i=1}^n EU_{ij}^2\right)^{1/2}}{\left(\frac{1}{n}\sum_{i=1}^n E|U_{ij}^3|\right)^{1/3}}.$$

Appendix E. Results for Section A.1

Proof of Lemma 1. Let $\delta=\widehat{\eta}_u-\eta_u$ and define

$$\widehat{R}(\eta)=\mathbb{E}_n[\rho_u(\tilde{y}_i-\tilde{x}_i'\eta)]-\mathbb{E}_n[\rho_u(\tilde{y}_i-\tilde{x}_i'\eta_u-r_{ui})]-\mathbb{E}_n[(u-1\{\tilde{y}_i\leqslant\tilde{x}_i'\eta_u+r_{ui}\})(\tilde{x}_i'\eta-\tilde{x}_i'\eta_u-r_{ui})].$$

By Lemma 14, $\widehat{R}(\eta)\geqslant 0$, $\bar{\mathrm{E}}[\widehat{R}(\eta_u)]\leqslant\bar{f}\|r_{ui}\|_{2,n}^2/2$ and with probability at least $1-\gamma$, $\widehat{R}(\eta_u)\leqslant\bar{R}_\gamma:=4\max\{\bar{f}\|r_{ui}\|_{2,n}^2,\ \|r_{ui}\|_{2,n}\sqrt{\log(8/\gamma)/n}\}\leqslant 4Cs\log(p/\gamma)/n$. By definition of $\widehat{\eta}_u$ we have

$$\begin{aligned}\widehat{R}(\widehat{\eta}_u)-\widehat{R}(\eta_u)+\mathbb{E}_n[(u-1\{\tilde{y}_i\leqslant\tilde{x}_i'\eta_u+r_{ui}\})\tilde{x}_i']\delta &= \mathbb{E}_n[\rho_u(\tilde{y}_i-\tilde{x}_i'\widehat{\eta}_u)]-\mathbb{E}_n[\rho_u(\tilde{y}_i-\tilde{x}_i'\eta_u)]\\ &\leqslant\tfrac{\lambda_u}{n}\|\eta_u\|_1-\tfrac{\lambda_u}{n}\|\widehat{\eta}_u\|_1.\end{aligned}\tag{E.55}$$

Let $N=\sqrt{8\mathbf{c}\bar{R}_\gamma/\underline{f}}+\frac{10}{\underline{f}}\left\{\bar{f}\|r_{ui}\|_{2,n}+\frac{3\mathbf{c}\lambda_u\sqrt{s}}{n\kappa_{2\mathbf{c}}}+\frac{8(1+2\mathbf{c})\sqrt{s\log(16p/\gamma)}}{\sqrt{n}\kappa_{2\mathbf{c}}}+\frac{8\mathbf{c}\sqrt{n}\bar{R}_\gamma\sqrt{\log(16p/\gamma)}}{\lambda_u\{s\log(p/\gamma)/n\}^{1/2}}\right\}$ denote the upper bound in the rate of convergence. Note that $N\geqslant\{s\log(p/\gamma)/n\}^{1/2}$. Suppose that the result is violated, so that $\|\tilde{x}_i'\delta\|_{2,n}>N$. Then by convexity of the objective function in (A.32), there is also a vector $\tilde{\delta}$ such that $\|\tilde{x}_i'\tilde{\delta}\|_{2,n}=N$, and

$$\mathbb{E}_n[\rho_u(\tilde{y}_i-\tilde{x}_i'(\tilde{\delta}+\eta_u))]-\mathbb{E}_n[\rho_u(\tilde{y}_i-\tilde{x}_i'\eta_u)]\ \leqslant\tfrac{\lambda_u}{n}\|\eta_u\|_1-\tfrac{\lambda_u}{n}\|\tilde{\delta}+\eta_u\|_1.\tag{E.56}$$

Next we will show that with high probability such $\tilde{\delta}$ cannot exist implying that $\|\tilde{x}_i'\delta\|_{2,n}\leqslant N$.

By the choice of $\lambda_u \geqslant c\Lambda_u(1-\gamma \mid \tilde{x})$ the event $\Omega_1 := \{\frac{\lambda_u}{n} \geqslant c\|\mathbb{E}_n[(u - 1\{\tilde{y}_i \leqslant \tilde{x}_i'\eta_u + r_{ui}\})\tilde{x}_i]\|_\infty\}$ occurs with probability at least $1-\gamma$. The event $\Omega_2 := \{\widehat{R}_1(\eta_u) \leqslant \bar{R}_\gamma\}$ also holds with probability at least $1-\gamma$. Under $\Omega_1 \cap \Omega_2$, and since $\widehat{R}(\eta) \geqslant 0$, we have

$$\begin{aligned} -\widehat{R}(\eta_u) - \tfrac{\lambda_u}{cn}\|\tilde{\delta}\|_1 &\leqslant \widehat{R}(\eta_u + \tilde{\eta}) - \widehat{R}(\eta_u) + \mathbb{E}_n[(u - 1\{\tilde{y}_i \leqslant \tilde{x}_i'\eta_u + r_{ui}\})\tilde{x}_i']\tilde{\delta} \\ &= \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'(\tilde{\delta} + \eta_u))] - \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \\ &\leqslant \tfrac{\lambda_u}{n}\|\eta_u\|_1 - \tfrac{\lambda_u}{n}\|\tilde{\delta} + \eta_u\|_1 \end{aligned} \tag{E.57}$$

so that for $\mathbf{c} = (c+1)/(c-1)$

$$\|\tilde{\delta}_{T_u^c}\|_1 \leqslant \mathbf{c}\|\tilde{\delta}_{T_u}\|_1 + \frac{nc}{\lambda_u(c-1)}\widehat{R}(\eta_u).$$

To establish that $\tilde{\delta} \in A_u := \Delta_{2\mathbf{c}} \cup \{v : \|\tilde{x}_i'v\|_{2,n} = N, \|v\|_1 \leqslant 2\mathbf{c}n\bar{R}_\gamma/\lambda_u\}$ we consider two cases. If $\|\tilde{\delta}_{T_u^c}\|_1 \geqslant 2\mathbf{c}\|\tilde{\delta}_{T_u}\|_1$ we have

$$\frac{1}{2}\|\tilde{\delta}_{T_u^c}\|_1 \leqslant \frac{nc}{\lambda_u(c-1)}\widehat{R}(\eta_u)$$

and consequentially

$$\|\tilde{\delta}\|_1 \leqslant \{1 + 1/(2c)\}\|\tilde{\delta}_{T_u^c}\|_1 \leqslant \frac{2n\mathbf{c}}{\lambda_u}\widehat{R}(\eta_u).$$

Otherwise $\|\tilde{\delta}_{T_u^c}\|_1 \leqslant 2\mathbf{c}\|\tilde{\delta}_{T_u}\|_1$, and we have

$$\|\tilde{\delta}\|_1 \leqslant (1+2\mathbf{c})\|\tilde{\delta}_{T_u}\|_1 \leqslant (1+2\mathbf{c})\sqrt{s}\|\tilde{x}_i'\tilde{\delta}\|_{2,n}/\kappa_{2\mathbf{c}}.$$

Thus with probability $1-2\gamma$, $\tilde{\delta} \in A_u$.

Therefore, under $\Omega_1 \cap \Omega_2$, from (E.56), applying Lemma 16 (part (1) and (3) to cover $\tilde{\delta} \in A_u$), for $\|\tilde{x}_i'\tilde{\delta}\|_{2,n} = N$ with probability at least $1-4\gamma$ we have

$$\begin{aligned} \bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'(\tilde{\delta} + \eta_u))] - \bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] &\leqslant \tfrac{\lambda_u}{n}\|\tilde{\delta}\|_1 + \tfrac{\|\tilde{x}_i'\tilde{\delta}\|_{2,n}}{\sqrt{n}}\left\{\tfrac{8(1+2\mathbf{c})\sqrt{s}}{\kappa_{2\mathbf{c}}} + \tfrac{8\mathbf{c}n\bar{R}_\gamma}{\lambda_u \underline{N}}\right\}\sqrt{\log(16p/\gamma)} \\ &\leqslant 2\mathbf{c}\bar{R}_\gamma + \|\tilde{x}_i'\tilde{\delta}\|_{2,n}\left[\tfrac{3\mathbf{c}\lambda_u\sqrt{s}}{n\kappa_{2\mathbf{c}}} + \left\{\tfrac{8(1+2\mathbf{c})\sqrt{s}}{\kappa_{2\mathbf{c}}} + \tfrac{8\mathbf{c}n\bar{R}_\gamma}{\lambda_u N}\right\}\tfrac{\sqrt{\log(16p/\gamma)}}{\sqrt{n}}\right] \end{aligned}$$

where we used the bound for $\|\tilde{\delta}\|_1 \leqslant (1+2\mathbf{c})\sqrt{s}\|\tilde{x}_i'\tilde{\delta}\|_{2,n}/\kappa_{2\mathbf{c}} + \frac{2n\mathbf{c}}{\lambda_u}\bar{R}_\gamma$.

Using Lemma 13, since by assumption $\sup_{\bar{\delta}\in A_u}\frac{\mathbb{E}_n[|r_{ui}||\tilde{x}_i'\bar{\delta}|^2]}{\mathbb{E}_n[|\tilde{x}_i'\bar{\delta}|^2]} \to 0$, we have

$$\bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \tilde{\delta})) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \geqslant -\bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'\tilde{\delta}\|_{2,n} + \frac{\underline{f}\|\tilde{x}_i'\tilde{\delta}\|_{2,n}^2}{4} \wedge \bar{q}_{A_u}\underline{f}\|\tilde{x}_i'\tilde{\delta}\|_{2,n}$$

Note that $N < 4\bar{q}_{A_u}$ for n sufficiently large by the assumed side condition, so that the minimum on the right hand side is achieved for the quadratic part. Therefore we have

$$\frac{\underline{f}\|\tilde{x}_i'\tilde{\delta}\|_{2,n}^2}{4} \leqslant 2\mathbf{c}\bar{R}_\gamma + \|\tilde{x}_i'\tilde{\delta}\|_{2,n}\left\{\bar{f}\|r_{ui}\|_{2,n} + \frac{3\mathbf{c}\lambda_u\sqrt{s}}{n\kappa_{2\mathbf{c}}} + \frac{8(1+2\mathbf{c})\sqrt{s\log(16p/\gamma)}}{\sqrt{n}\kappa_{2\mathbf{c}}} + \frac{8\mathbf{c}\sqrt{n}\bar{R}_\gamma\sqrt{\log(16p/\gamma)}}{\lambda_u N}\right\}$$

which implies that

$$\|\tilde{x}_i'\tilde{\delta}\|_{2,n} \leqslant \sqrt{8\mathbf{c}\bar{R}_\gamma/\underline{f}} + \frac{8}{\underline{f}}\left\{\bar{f}\|r_{ui}\|_{2,n} + \frac{3\mathbf{c}\lambda_u\sqrt{s}}{n\kappa_{2\mathbf{c}}} + \frac{8(1+2\mathbf{c})\sqrt{s\log(16p/\gamma)}}{\sqrt{n}\kappa_{2\mathbf{c}}} + \frac{8\mathbf{c}\sqrt{n}\bar{R}_\gamma\sqrt{\log(16p/\gamma)}}{\lambda_u N}\right\}$$

which violates the assumed condition that $\|\tilde{x}_i'\tilde{\delta}\|_{2,n} = N$ since $N > \{s\log(p/\gamma)/n\}^{1/2}$. □

Proof of Lemma 2. Let $\widehat{\delta}_u = \widehat{\eta}_u - \eta_u$. By optimality of $\widetilde{\eta}_u$ in (A.32) we have with probability $1-\gamma$

$$\mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\widetilde{\eta}_u)] - \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \quad \leqslant \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\widehat{\eta}_u)] - \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \leqslant \widehat{Q}. \tag{E.58}$$

Let $N = 2\bar{f}\bar{r}_u + A_{\varepsilon,n} + 2\widehat{Q}^{1/2}$ denote the upper bound in the rate of convergence where $A_{\varepsilon,n}$ is defined below. Suppose that the result is violated, so that $\|\tilde{x}_i'(\widetilde{\eta}_u - \eta_u)\|_{2,n} > N$. Then by convexity of the objective function in (A.32), there is also a vector $\widetilde{\delta}_u$ such that $\|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n} = N$, $\|\widetilde{\delta}_u\|_0 = \|\widetilde{\eta}_u - \eta_u\|_0 \leqslant \widehat{s}_u + s$ and

$$\mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \widetilde{\delta}_u))] - \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \leqslant \widehat{Q}. \tag{E.59}$$

Next we will show that with high probability such $\widetilde{\delta}_u$ cannot exist implying that $\|\tilde{x}_i'(\widetilde{\eta}_u - \eta_u)\|_{2,n} \leqslant N$ with high probability.

By Lemma 16, with probability at least $1-\varepsilon$, we have

$$\frac{|(\mathbb{E}_n - \bar{\mathrm{E}})[\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \widetilde{\delta}_u)) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)]|}{\|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n}} \leqslant 8\sqrt{\frac{(\widehat{s}_u + s)\log(16p/\varepsilon)}{n\phi_{\min}(\widehat{s}_u + s)}} =: A_{\varepsilon,n}. \tag{E.60}$$

Thus combining relations (E.58) and (E.60), we have

$$\bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \widetilde{\delta}_u))] - \bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \quad \leqslant \quad \|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n} A_{\varepsilon,n} + \widehat{Q}$$

with probability at least $1-\varepsilon$. Invoking the sparse identifiability relation of Lemma 13, with the same probability, since $\sup_{\|\delta\|_0 \leqslant \widehat{s}_u + s} \frac{\mathbb{E}_n[|r_{ui}|\ |\tilde{x}_i'\theta|^2]}{\mathbb{E}_n[|\tilde{x}_i'\theta|^2]} \to 0$ by assumption,

$$(\underline{f}\|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n}^2/4) \wedge \left(\widetilde{q}_{\widehat{s}_u}\underline{f}\|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n}\right) \quad \leqslant \quad \|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n}\left\{\bar{f}\|r_{ui}\|_{2,n} + A_{\varepsilon,n}\right\} + \widehat{Q}.$$

where $\widetilde{q}_{\widehat{s}_u} := \frac{\underline{f}^{3/2}}{2\bar{f}'} \inf_{\|\delta\|_0 \leqslant \widehat{s}_u + s} \frac{\|\tilde{x}_i'\theta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\theta|^3]}$.

Under the assumed growth condition, we have $N < 4\widetilde{q}_{\widehat{s}_u}$ for n sufficiently large and the minimum is achieved in the quadratic part. Therefore, for n sufficiently large, we have

$$\|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n} \leqslant \bar{f}\|r_{ui}\|_{2,n} + A_{\varepsilon,n} + 2\widehat{Q}^{1/2} < N$$

Thus with probability at least $1 - \varepsilon - \gamma - o(1)$ we have $\|\tilde{x}_i'\widetilde{\delta}_u\|_{2,n} < N$ which contradicts its definition. Therefore, $\|\tilde{x}_i'(\widetilde{\eta}_u - \eta_u)\|_{2,n} \leqslant N$ with probability at least $1 - \gamma - \varepsilon - o(1)$. □

E.1. **Technical Lemmas for High-Dimensional Quantile Regression.**

Lemma 13. *For a subset $A \subset \mathbb{R}^p$ let*

$$\bar{q}_A = (1/2) \cdot (\underline{f}^{3/2}/\bar{f}') \cdot \inf_{\delta \in A} \mathbb{E}_n\left[|\tilde{x}_i'\delta|^2\right]^{3/2} / \mathbb{E}_n\left[|\tilde{x}_i'\delta|^3\right]$$

and assume that for all $\delta \in A$

$$\bar{\mathrm{E}}\left[|r_{ui}| \cdot |\tilde{x}_i'\delta|^2\right] \leqslant (\underline{f}/[4\bar{f}'])\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^2].$$

Then, we have

$$\bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \delta))] - \bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \geqslant \frac{\underline{f}\|\tilde{x}_i'\delta\|_{2,n}^2}{4} \wedge \left\{\bar{q}_A\underline{f}\|\tilde{x}_i'\delta\|_{2,n}\right\} - \bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'\delta\|_{2,n}.$$

Proof of Lemma 13. Let $T = \text{support}(\eta_u)$, $Q_u(\eta) := \bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta)]$, $J_u = (1/2)\mathbb{E}_n\left[f_i \tilde{x}_i \tilde{x}_i'\right]$ and define $\|\delta\|_u = \|J_u^{1/2}\delta\|$. The proof proceeds in steps.

Step 1. (Minoration). Define the maximal radius over which the criterion function can be minorated by a quadratic function

$$r_A = \sup_r \left\{ r \ : Q_u(\eta_u + \delta) - Q_u(\eta_u) + \bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'\delta\|_{2,n} \geqslant \frac{1}{2}\|\delta\|_u^2, \text{ for all } \delta \in A, \ \|\delta\|_u \leqslant r \right\}.$$

Step 2 below shows that $r_A \geqslant \bar{q}_A$. By construction of r_A and the convexity of $Q_u(\cdot)$ and $\|\cdot\|_u$,

$$\begin{array}{l} Q_u(\eta_u + \delta) - Q_u(\eta_u) + \bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'\delta\|_{2,n} \geqslant \\ \geqslant \frac{\|\delta\|_u^2}{2} \wedge \left\{ \frac{\|\delta\|_u}{r_A} \cdot \inf_{\tilde{\delta} \in A, \|\tilde{\delta}\|_u \geqslant r_A} Q_u(\eta_u + \tilde{\delta}) - Q_u(\eta_u) + \bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'\tilde{\delta}\|_{2,n} \right\} \\ \geqslant \frac{\|\delta\|_u^2}{2} \wedge \left\{ \frac{\|\delta\|_u}{r_A} \frac{r_A^2}{4} \right\} \geqslant \frac{\|\delta\|_u^2}{2} \wedge \left\{ \bar{q}_A \|\delta\|_u \right\}. \end{array}$$

Step 2. ($r_A \geqslant \bar{q}_A$) Let $F_{\tilde{y}|\tilde{x}}$ denote the conditional distribution of $\tilde{y}$ given $\tilde{x}$. From [20], for any two scalars w and v we have that

$$\rho_u(w - v) - \rho_u(w) = -v(u - 1\{w \leqslant 0\}) + \int_0^v (1\{w \leqslant z\} - 1\{w \leqslant 0\})dz. \tag{E.61}$$

We will use (E.61) with $w = \tilde{y}_i - \tilde{x}_i'\eta_u$ and $v = \tilde{x}_i'\delta$. Using the law of iterated expectations and mean value expansion, we obtain for $\tilde{t}_{\tilde{x}_i,t} \in [0,t]$

$$\begin{array}{l} Q_u(\eta_u + \delta) - Q_u(\eta_u) + \bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'\delta\|_{2,n} \geqslant \\ Q_u(\eta_u + \delta) - Q_u(\eta_u) + \bar{\mathrm{E}}\left[(u - 1\{\tilde{y}_i \leqslant \tilde{x}_i'\eta_u\})\tilde{x}_i'\delta\right] = \\ = \bar{\mathrm{E}}\left[\int_0^{\tilde{x}_i'\delta} F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + t) - F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u)dt\right] \\ = \bar{\mathrm{E}}\left[\int_0^{\tilde{x}_i'\delta} t f_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u) + \frac{t^2}{2} f'_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + \tilde{t}_{\tilde{x},t})dt\right] \\ \geqslant \|\delta\|_u^2 - \frac{1}{6}\bar{f}'\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^3] - \bar{\mathrm{E}}\left[\int_0^{\tilde{x}_i'\delta} t[f_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u) - f_{\tilde{y}_i|\tilde{x}_i}(g_{ui})]dt\right] \\ \geqslant \frac{1}{2}\|\delta\|_u^2 + \frac{1}{4}\underline{f}\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^2] - \frac{1}{6}\bar{f}'\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^3] - (\bar{f}'/2)\bar{\mathrm{E}}\left[|\tilde{x}_i'\eta_u - g_{ui}| \cdot |\tilde{x}_i'\delta|^2\right]. \end{array} \tag{E.62}$$

where the first inequality follows noting that $F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + r_{ui}) = u$ and $|F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + r_{ui}) - F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u)| \leqslant \bar{f}|r_{ui}|$.

Moreover, by assumption we have

$$\begin{array}{ll} \bar{\mathrm{E}}\left[|\tilde{x}_i'\eta_u - g_{ui}| \cdot |\tilde{x}_i'\delta|^2\right] & = \bar{\mathrm{E}}\left[|r_{ui}| \cdot |\tilde{x}_i'\delta|^2\right] \\ & \leqslant (\underline{f}/8)(2/\bar{f}')\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^2] \end{array} \tag{E.63}$$

Note that for any δ such that $\|\delta\|_u \leqslant \bar{q}_A$ we have $\|\delta\|_u \leqslant \bar{q}_A \leqslant (1/2)\cdot(\underline{f}^{3/2}/\bar{f}')\cdot\bar{\mathrm{E}}\left[|\tilde{x}_i'\delta|^2\right]^{3/2}/\bar{\mathrm{E}}\left[|\tilde{x}_i'\delta|^3\right]$, it follows that $(1/6)\bar{f}'\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^3] \leqslant (1/8)\underline{f}\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^2]$. Combining this with (E.63) we have

$$\frac{1}{4}\underline{f}\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^2] - \frac{1}{6}\bar{f}'\bar{\mathrm{E}}[|\tilde{x}_i'\delta|^3] - (\bar{f}'/2)\bar{\mathrm{E}}\left[|\tilde{x}_i'\eta_u - g_{ui}| \cdot |\tilde{x}_i'\delta|^2\right] \geqslant 0. \tag{E.64}$$

Combining (E.62) and (E.64) we have $r_A \geqslant \bar{q}_A$. □

Lemma 14. *Under Condition PQR we have* $\bar{\mathrm{E}}[\widehat{R}(\eta_u)] \leqslant \bar{f}\|r_{ui}\|_{2,n}^2/2$, $\widehat{R}(\eta_u) \geqslant 0$ *and*

$$P(\widehat{R}(\eta_u) \geqslant 4\max\{\bar{f}\|r_{ui}\|_{2,n}^2, \|r_{ui}\|_{2,n}\sqrt{\log(8/\gamma)/n}\}) \leqslant \gamma.$$

Proof of Lemma 14. We have that $\widehat{R}(\eta_u) \geqslant 0$ by convexity of ρ_u. Let $\epsilon_{ui} = \tilde{y}_i - \tilde{x}_i'\eta_u - r_{ui}$. By Knight's identity, $\widehat{R}(\eta_u) = -\mathbb{E}_n[r_{ui}\int_0^1 1\{\epsilon_{ui} \leqslant -tr_{ui}\} - 1\{\epsilon_{ui} \leqslant 0\}\, dt] \geqslant 0$.

$$\begin{aligned}\bar{\mathrm{E}}[\widehat{R}(\eta_u)] &= \mathbb{E}_n[r_{ui}\int_0^1 F_{y_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + (1-t)r_{ui}) - F_{y_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + r_{ui})\, dt] \\ &\leqslant \mathbb{E}_n[r_{ui}\int_0^1 \bar{f}tr_{ui}dt] \leqslant \bar{f}\|r_{ui}\|_{2,n}^2/2.\end{aligned}$$

Therefore $P(\widehat{R}(\eta_u) \leqslant 2\bar{f}\|r_{ui}\|_{2,n}^2) \geqslant 1/2$ by Markov's inequality.

Define $z_{ui} := -\int_0^1 1\{\epsilon_{ui} \leqslant -tr_{ui}\} - 1\{\epsilon_{ui} \leqslant 0\}\, dt$, so that $\widehat{R}(\eta_u) = \mathbb{E}_n[r_{ui}z_{ui}]$. We have $P(\mathbb{E}_n[r_{ui}z_{ui}] \leqslant 2\bar{f}\|r_{ui}\|_{2,n}^2) \geqslant 1/2$ so that for $t \geqslant 4\bar{f}\|r_{ui}\|_{2,n}^2$ we have by Lemma 2.3.7 in [38]

$$\frac{1}{2}P(|\mathbb{E}_n[r_{ui}z_{ui}]| \geqslant t) \leqslant 2P(|\mathbb{E}_n[r_{ui}z_{ui}\epsilon_i]| > t/4)$$

Since the $r_{ui}z_{ui}\epsilon_i$ is a symmetric random variable and $|z_{ui}| \leqslant 1$, by Theorem 2.15 in [16] we have

$$P(\sqrt{n}|\mathbb{E}_n[r_{ui}z_{ui}\epsilon_i]| > \bar{t}\sqrt{\mathbb{E}_n[r_{ui}^2]}) \leqslant P(\sqrt{n}|\mathbb{E}_n[r_{ui}z_{ui}\epsilon_i]| > \bar{t}\sqrt{\mathbb{E}_n[r_{ui}^2z_{ui}^2]}) \leqslant 2\exp(-\bar{t}^2/2) \leqslant \gamma/8$$

for $\bar{t} \geqslant \sqrt{2\log(8/\gamma)}$. Setting $t = 4\max\{\bar{f}\|r_{ui}\|_{2,n}^2, \|r_{ui}\|_{2,n}\sqrt{\log(8/\gamma)/n}\}$ we have

$$P(\mathbb{E}_n[r_{ui}z_{ui}] \geqslant t) \leqslant 4P(\mathbb{E}_n[r_{ui}z_{ui}\epsilon_i] > t/4) \leqslant \gamma.$$

□

Lemma 15. *Under Condition PQR, for $\|\widehat{\eta}_u\|_0 \leqslant k$, $\underline{N} \leqslant \|\tilde{x}_i'(\widehat{\eta}_u - \eta_u)\|_{2,n} \leqslant \bar{N}$, we have with probability $1-\gamma$*

$$\mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\widehat{\eta}_u)] - \mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \leqslant \frac{\|\tilde{x}_i'(\widehat{\eta}_u - \eta_u)\|_{2,n}}{\sqrt{n}}\left\{4 + 4\sqrt{\frac{(k+s)\log(16p\{1+3\sqrt{n}\log(\frac{\bar{N}}{\underline{N}})\}/\gamma)}{\phi_{\min}(k+s)}}\right\}$$

$$+\bar{f}\|\tilde{x}_i'(\widehat{\eta}_u - \eta_u)\|_{2,n}^2 + \bar{f}\|r_{ui}\|_{2,n}\|\tilde{x}_i'(\widehat{\eta}_u - \eta_u)\|_{2,n}.$$

Proof of Lemma 15. It follows from

$$\mathbb{E}_n[\rho_u(\tilde{y}_i - \tilde{x}_i'\widehat{\eta}_u) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)] \leqslant |(\mathbb{E}_n - \bar{\mathrm{E}})[\rho_u(\tilde{y}_i - \tilde{x}_i'\widehat{\eta}_u)\rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)]| + \bar{\mathrm{E}}[\rho_u(\tilde{y}_i - \tilde{x}_i'\widehat{\eta}_u) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)]$$

where the first term is bounded by Lemma 16 and the second term is bounded by (E.62) noting that

$$\mathbb{E}_n\left[\int_0^{\tilde{x}_i'\delta} F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u + t) - F_{\tilde{y}_i|\tilde{x}_i}(\tilde{x}_i'\eta_u)dt\right] \leqslant \bar{f}\mathbb{E}_n\left[\int_0^{\tilde{x}_i'\delta} t dt\right] \leqslant \bar{f}\|\tilde{x}_i'\delta\|_{2,n}^2.$$

□

Lemma 16. *Conditional on $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ we have with probability $1-\gamma$, for vectors in the restricted set*

$$\sup_{\substack{\delta \in \Delta_{\mathbf{c}}, \\ \underline{N} \leqslant \|\tilde{x}_i'\delta\|_{2,n} \leqslant \bar{N}}} \left|\mathbb{G}_n\left(\frac{\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \delta)) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)}{\|\tilde{x}_i'\delta\|_{2,n}}\right)\right| \leqslant 4 + \frac{4(1+\mathbf{c})\sqrt{s\log(16p\{1+3\sqrt{n}\log(\frac{\bar{N}}{\underline{N}})\}/\gamma)}}{\kappa_{\mathbf{c}}}$$

Similarly, for sparse vectors

$$\sup_{\substack{1 \leqslant \|\delta\|_0 \leqslant k, \\ \underline{N} \leqslant \|\tilde{x}_i'\delta\|_{2,n} \leqslant \bar{N}}} \left| \mathbb{G}_n \left(\frac{\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \delta)) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)}{\|\tilde{x}_i'\delta\|_{2,n}} \right) \right| \leqslant 4 + 4\sqrt{\frac{k \log(16p\{1 + 3\sqrt{n}\log(\bar{N}/\underline{N})\}/\gamma)}{\phi_{\min}(k)}}$$

Similarly, for ℓ_1-bounded vectors

$$\sup_{\substack{\|\delta\|_1 \leqslant R_1, \\ \underline{N} \leqslant \|\tilde{x}_i'\delta\|_{2,n} \leqslant \bar{N}}} \left| \mathbb{G}_n \left(\frac{\rho_u(\tilde{y}_i - \tilde{x}_i'(\eta_u + \delta)) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u)}{\|\tilde{x}_i'\delta\|_{2,n}} \right) \right| \leqslant 4 + 4\frac{R_1}{\underline{N}}\sqrt{\log(16p\{1 + 3\sqrt{n}\log(\bar{N}/\underline{N})\}/\gamma)}$$

Proof of Lemma 16. Let $w_i(b) = \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u + b) - \rho_u(\tilde{y}_i - \tilde{x}_i'\eta_u) \leqslant |b|$. Note that $w_i(b) - w_i(a) \leqslant |b - a|$.

For any $\delta \in \mathbb{R}^p$, since ρ_u is 1-Lipschitz, we have

$$\operatorname{var}\left(\mathbb{G}_n \left(\frac{w_i(\tilde{x}_i'\delta)}{\|\tilde{x}_i'\delta\|_{2,n}} \right) \right) \quad \leqslant \frac{\mathbb{E}_n[\{w_i(\tilde{x}_i'\delta)\}^2]}{\|\tilde{x}_i'\delta\|_{2,n}^2} \leqslant \frac{\mathbb{E}_n[|\tilde{x}_i'\delta|^2]}{\|\tilde{x}_i'\delta\|_{2,n}^2} \leqslant 1.$$

Then, by Lemma 2.3.7 in [37] (Symmetrization for Probabilities) we have for any $M > 1$

$$P\left(\sup_{\delta \in \Delta_{\mathbf{c}}} \left| \mathbb{G}_n \left(\frac{w_i(\tilde{x}_i'\delta)}{\|\tilde{x}_i'\delta\|_{2,n}} \right) \right| \geqslant M \right) \leqslant \frac{2}{1 - M^{-2}} P\left(\sup_{\delta \in \Delta_{\mathbf{c}}} \left| \mathbb{G}_n^o \left(\frac{w_i(\tilde{x}_i'\delta)}{\|\tilde{x}_i'\delta\|_{2,n}} \right) \right| \geqslant M/4 \right)$$

where $\mathbb{G}_n^o$ is the symmetrized process.

Consider $\mathcal{F}_t = \{\delta \in \Delta_{\mathbf{c}} : \|\tilde{x}_i'\delta\|_{2,n} = t\}$. We will consider the families of $\mathcal{F}_t$ for $t \in [\underline{N}, \bar{N}]$. For any $\delta \in \mathcal{F}_t$, $t \leqslant \tilde{t}$ we have

$$\begin{aligned}
\left| \mathbb{G}_n^o \left(\tfrac{w_i(\tilde{x}_i'\delta)}{t} - \tfrac{w_i(\tilde{x}_i'\delta(\tilde{t}/t))}{\tilde{t}} \right) \right| &\leqslant \left| \mathbb{G}_n^o \left(\tfrac{w_i(\tilde{x}_i'\delta)}{t} - \tfrac{w_i(\tilde{x}_i'\delta(\tilde{t}/t))}{t} \right) \right| + \left| \mathbb{G}_n^o \left(\tfrac{w_i(\tilde{x}_i'\delta(\tilde{t}/t))}{t} - \tfrac{w_i(\tilde{x}_i'\delta(\tilde{t}/t))}{\tilde{t}} \right) \right| \\
&= \tfrac{1}{t} \left| \mathbb{G}_n^o \left(w_i(\tilde{x}_i'\delta) - w_i(\tilde{x}_i'\delta[\tilde{t}/t]) \right) \right| + \left| \mathbb{G}_n^o \left(w_i(\tilde{x}_i'\delta(\tilde{t}/t)) \right) \right| \cdot \left| \tfrac{1}{t} - \tfrac{1}{\tilde{t}} \right| \\
&\leqslant \sqrt{n}\mathbb{E}_n \left(\tfrac{|\tilde{x}_i'\delta|}{t} \right) \tfrac{|t - \tilde{t}|}{t} + \sqrt{n}\mathbb{E}_n \left(|\tilde{x}_i'\delta| \right) \tfrac{\tilde{t}}{t} \left| \tfrac{1}{t} - \tfrac{1}{\tilde{t}} \right| \\
&= 2\sqrt{n}\mathbb{E}_n \left(\tfrac{|\tilde{x}_i'\delta|}{t} \right) \left| \tfrac{t - \tilde{t}}{t} \right| \leqslant 2\sqrt{n} \left| \tfrac{t - \tilde{t}}{t} \right|.
\end{aligned}$$

Let $\mathcal{T}$ be a ε-net $\{\underline{N} =: t_1, t_2, \ldots, t_K := \bar{N}\}$ of $[\underline{N}, \bar{N}]$ such that $|t_k - t_{k+1}|/t_k \leqslant 1/[2\sqrt{n}]$. Note that we can achieve that with $|\mathcal{T}| \leqslant 3\sqrt{n}\log(\bar{N}/\underline{N})$.

Therefore we have

$$\sup_{\delta \in \Delta_{\mathbf{c}}} \left| \mathbb{G}_n^o \left(\frac{w_i(\tilde{x}_i'\delta)}{\|\tilde{x}_i'\delta\|_{2,n}} \right) \right| \leqslant 1 + \sup_{t \in \mathcal{T}} \sup_{\delta \in \Delta_{\mathbf{c}}, \|\tilde{x}_i'\delta\|_{2,n} = t} \left| \mathbb{G}_n^o \left(\frac{w_i(\tilde{x}_i'\delta)}{t} \right) \right| =: 1 + \mathcal{A}^o.$$

$$\begin{aligned}
P(\mathcal{A}^o \geqslant K) &\leqslant \min_{\psi \geqslant 0} \exp(-\psi K) \mathrm{E}[\exp(\psi \mathcal{A}^o)] \\
&\leqslant 8p|\mathcal{T}| \min_{\psi \geqslant 0} \exp(-\psi K) \exp\left(8\psi^2 \tfrac{s(1+\mathbf{c})^2}{\kappa_{\mathbf{c}}^2} \right) \\
&\leqslant 8p|\mathcal{T}| \exp(-K^2/[16\tfrac{s(1+\mathbf{c})^2}{\kappa_{\mathbf{c}}^2}])
\end{aligned}$$

where we set $\psi = K/[16\frac{s(1+\mathbf{c})^2}{\kappa_{\mathbf{c}}^2}]$ and bounded

$$\begin{aligned}
\mathrm{E}\left[\exp\left(\psi\mathcal{A}^o\right)\right] &\leqslant_{(1)} 2|\mathcal{T}|\sup_{t\in\mathcal{T}}\mathrm{E}\left[\exp\left(\psi\sup_{\delta\in\Delta_{\mathbf{c}},\|\tilde{x}_i'\delta\|_{2,n}=t}\mathbb{G}_n^o\left(\frac{w_i(\tilde{x}_i'\delta)}{t}\right)\right)\right]\\
&\leqslant_{(2)} 2|\mathcal{T}|\sup_{t\in\mathcal{T}}\mathrm{E}\left[\exp\left(2\psi\sup_{\delta\in\Delta_{\mathbf{c}},\|\tilde{x}_i'\delta\|_{2,n}=t}\mathbb{G}_n^o\left(\frac{\tilde{x}_i'\delta}{t}\right)\right)\right]\\
&\leqslant_{(3)} 2|\mathcal{T}|\sup_{t\in\mathcal{T}}\mathrm{E}\left[\exp\left(2\psi\left[\sup_{\delta\in\Delta_{\mathbf{c}},\|\tilde{x}_i'\delta\|_{2,n}=t}2\frac{\|\delta\|_1}{t}\right]\max_{j\leqslant p}|\mathbb{G}_n^o(\tilde{x}_{ij})|\right)\right]\\
&\leqslant_{(4)} 2|\mathcal{T}|\mathrm{E}\left[\exp\left(4\psi\frac{\sqrt{s}(1+\mathbf{c})}{\kappa_{\mathbf{c}}}\max_{j\leqslant p}|\mathbb{G}_n^o(\tilde{x}_{ij})|\right)\right]\\
&\leqslant_{(5)} 4p|\mathcal{T}|\max_{j\leqslant p}\mathrm{E}\left[\exp\left(4\psi\frac{\sqrt{s}(1+\mathbf{c})}{\kappa_{\mathbf{c}}}\mathbb{G}_n^o(\tilde{x}_{ij})\right)\right]\\
&\leqslant_{(6)} 8p|\mathcal{T}|\exp\left(8\psi^2\ \frac{s(1+\mathbf{c})^2}{\kappa_{\mathbf{c}}^2}\right)
\end{aligned}$$

where (1) follows by $\exp(\max_{i\in I}|z_i|)\leqslant 2|I|\max_{i\in I}\exp(z_i)$, (2) by contraction principle (Theorem 4.12 [24]), (3) $|\mathbb{G}_n^o(\tilde{x}_i'\delta)|\leqslant\|\delta\|_1\|\mathbb{G}_n^o(\tilde{x}_i)\|_\infty$, (4) $\sqrt{s}(1+\mathbf{c})\|\tilde{x}_i'\delta\|_{2,n}/\|\delta\|_1\geqslant\kappa_{\mathbf{c}}$, (6) $\mathbb{E}_n[x_{ij}^2]=1$ and $\exp(z)+\exp(-z)\leqslant 2\exp(z^2/2)$.

The second result follows similarly by noting that

$$\sup_{1\leqslant\|\delta\|_0\leqslant k,\|\tilde{x}_i'\delta\|_{2,n}=t}\frac{\|\delta\|_1}{t}\leqslant\sup_{1\leqslant\|\delta\|_0\leqslant k,\|\tilde{x}_i'\delta\|_{2,n}=t}\frac{\sqrt{k}\|\tilde{x}_i'\delta\|_{2,n}}{t\sqrt{\phi_{\min}(k)}}=\frac{\sqrt{k}}{\sqrt{\phi_{\min}(k)}}.$$

The third result follows similarly by noting that for ant $t\in[\underline{N},\bar{N}]$

$$\sup_{\|\delta\|_1\leqslant R_1,\|\tilde{x}_i'\delta\|_{2,n}=t}\frac{\|\delta\|_1}{t}\leqslant\frac{R_1}{\underline{N}}.$$

□

Appendix F. Results for Section A.2

Lemma 17 (Choice of λ)**.** *Suppose Condition WL holds, let $c'>c>1$, $\gamma=1/(n\vee p)$, and $\lambda=2c'\sqrt{n}\Phi^{-1}(1-\gamma/2p)$. Then for $n\geqslant n_0(\delta_n,c',c)$ large enough*

$$P(\lambda/n\geqslant 2c\|\widehat{\Gamma}_{\tau 0}^{-1}\mathbb{E}_n[f_ix_iv_i]\|_\infty)\geqslant 1-\gamma\{1+o(1)\}+4\Delta_n.$$

Proof of Lemma 17. Since $\widehat{\Gamma}_{\tau 0jj}=\sqrt{\mathbb{E}_n[\widehat{f}_i^2x_{ij}^2v_i^2]}$ and $\Gamma_{\tau 0jj}=\sqrt{\mathbb{E}_n[f_i^2x_{ij}^2v_i^2]}$, with probability at least $1-\Delta_n$ we have

$$\max_{j\leqslant p}|\widehat{\Gamma}_{\tau 0jj}-\Gamma_{\tau 0jj}|\leqslant\max_{j\leqslant p}\sqrt{\mathbb{E}_n[(\widehat{f}_i-f_i)^2x_{ij}^2v_i^2]}\leqslant\delta_n^{1/2}$$

by Condition WL(iii). Further, Condition WL implies that $\Gamma_{\tau 0jj}$ is bounded away from zero and from above uniformly in $j=1,\ldots,p$ and n. Thus we have $\|\widehat{\Gamma}_{\tau 0}^{-1}\Gamma_{\tau 0}\|_\infty\to_P 1$, so that $\|\widehat{\Gamma}_{\tau 0}^{-1}\Gamma_{\tau 0}\|_\infty\leqslant\sqrt{c'/c}$ with probability $1-\Delta_n$ for $n\geqslant n_0(\delta_n,c',c,\Gamma_{\tau 0})$. By the triangle inequality

$$\|\widehat{\Gamma}_{\tau 0}^{-1}\mathbb{E}_n[f_ix_iv_i]\|_\infty\leqslant\|\widehat{\Gamma}_{\tau 0}^{-1}\Gamma_{\tau 0}\|_\infty\|\Gamma_{\tau 0}^{-1}\mathbb{E}_n[f_ix_iv_i]\|_\infty \tag{F.65}$$

Using Lemma 12, based on self-normalized moderate deviation theory, we have

$$P\left(\max_{j\leqslant p}\left|\frac{\sqrt{n}\mathbb{E}_n[f_i x_{ij} v_i]}{\sqrt{\mathbb{E}_n[f_i^2 x_{ij}^2 v_i^2]}}\right| > \Phi^{-1}(1-\gamma/2p)\right) \leqslant 2p\Phi(\Phi^{-1}(1-\gamma/2p))(1+o(1)) \leqslant \gamma\{1+o(1)\}$$

by Condition WL. □

Proof of Lemma 3. Let $\widehat{\delta} = \widehat{\theta}_\tau - \theta_\tau$. By definition of $\widehat{\theta}_\tau$ we have

$$\begin{aligned}\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2] - 2\mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta_\tau)x_i]'\widehat{\delta} &= \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\widehat{\theta}_\tau)^2] - \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta_\tau)^2] \\ &\leqslant \tfrac{\lambda}{n}\|\widehat{\Gamma}_\tau\theta_\tau\|_1 - \tfrac{\lambda}{n}\|\widehat{\Gamma}_\tau\widehat{\theta}_\tau\|_1 \leqslant \tfrac{\lambda}{n}\|\widehat{\Gamma}_\tau\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\|\widehat{\Gamma}_\tau\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \\ &\leqslant \tfrac{\lambda}{n}u\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\ell\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1\end{aligned} \tag{F.66}$$

Therefore, using that $c_f^2 = \mathbb{E}_n[(\widehat{f}_i^2 - f_i^2)^2 v_i^2/f_i^2]$ and $c_r^2 = \mathbb{E}_n[\widehat{f}^2 r_{\theta\tau i}^2]$, we have

$$\begin{aligned}\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2] &\leqslant 2\mathbb{E}_n[(\widehat{f}_i^2 - f_i^2)v_i x_i/f_i]'\widehat{\delta} + 2\mathbb{E}_n[\widehat{f}_i^2 r_{\theta\tau i}x_i]'\widehat{\delta} + 2(\widehat{\Gamma}_0^{-1}\mathbb{E}_n[f_i v_i x_i])'(\widehat{\Gamma}_{\tau 0}\widehat{\delta}) + \tfrac{\lambda}{n}u\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\ell\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \\ &\leqslant 2\{c_f + c_r\}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} + 2\|\widehat{\Gamma}_0^{-1}\mathbb{E}_n[f_i^2(d_i - x_i'\theta_\tau)x_i]\|_\infty\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}\|_1 + \tfrac{\lambda}{n}u\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\ell\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \\ &\leqslant 2\{c_f + c_r\}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} + \tfrac{\lambda}{cn}\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}\|_1 + \tfrac{\lambda}{n}u\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\ell\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \\ &\leqslant 2\{c_f + c_r\}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} + \tfrac{\lambda}{n}\left(u + \tfrac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\left(\ell - \tfrac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1\end{aligned} \tag{F.67}$$

Let $\tilde{\mathbf{c}} = \frac{cu+1}{c\ell-1}\|\widehat{\Gamma}_{\tau 0}\|_\infty\|\widehat{\Gamma}_{\tau 0}^{-1}\|_\infty$. If $\widehat{\delta} \notin \Delta_{\tilde{\mathbf{c}}}$ we have $\left(u + \frac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 \leqslant \left(\ell - \frac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1$ so that

$$\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} \leqslant 2\{c_f + c_r\}.$$

Otherwise assume $\widehat{\delta} \in \Delta_{\tilde{\mathbf{c}}}$. In this case (F.67) yields

$$\begin{aligned}\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2] &\leqslant 2\{c_f + c_r\}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} + \tfrac{\lambda}{n}\left(u + \tfrac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 - \tfrac{\lambda}{n}\left(\ell - \tfrac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \\ &\leqslant 2\{c_f + c_r\}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} + \tfrac{\lambda}{n}\left(u + \tfrac{1}{c}\right)\sqrt{s}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2}/\widehat{\kappa}_{\tilde{\mathbf{c}}}\end{aligned}$$

which implies

$$\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2} \leqslant 2\{c_f + c_r\} + \frac{\lambda\sqrt{s}}{n\widehat{\kappa}_{\tilde{\mathbf{c}}}}\left(u + \frac{1}{c}\right)$$

To establish the ℓ_1-bound, first assume that $\widehat{\delta} \in \Delta_{2\tilde{\mathbf{c}}}$. In that case

$$\|\widehat{\delta}\|_1 \leqslant (1 + 2\tilde{\mathbf{c}})\|\widehat{\delta}_{T_{\theta\tau}}\|_1 \leqslant \sqrt{s}\{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2}/\widehat{\kappa}_{2\tilde{\mathbf{c}}} \leqslant 2\frac{\sqrt{s}\{c_f + c_r\}}{\widehat{\kappa}_{2\tilde{\mathbf{c}}}} + \frac{\lambda s}{n\widehat{\kappa}_{\tilde{\mathbf{c}}}\widehat{\kappa}_{2\tilde{\mathbf{c}}}}\left(u + \frac{1}{c}\right).$$

Otherwise note that $\widehat{\delta} \notin \Delta_{2\tilde{\mathbf{c}}}$ implies that $\left(u + \frac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}}\|_1 \leqslant \frac{1}{2}\cdot\left(\ell - \frac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1$ so that (F.67) gives

$$\frac{1}{2}\frac{\lambda}{n}\cdot\left(\ell - \frac{1}{c}\right)\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \leqslant \{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2}\left(2\{\widehat{c}_f + \widehat{c}_r\} - \{\mathbb{E}_n[\widehat{f}_i^2(x_i'\widehat{\delta})^2]\}^{1/2}\right) \leqslant \{\widehat{c}_f + \widehat{c}_r\}^2.$$

Therefore

$$\|\widehat{\delta}\|_1 \leqslant \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\|\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \leqslant \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\|\widehat{\Gamma}_{\tau 0}^{-1}\|_\infty\|\widehat{\Gamma}_{\tau 0}\widehat{\delta}_{T_{\theta\tau}^c}\|_1 \leqslant \left(1 + \frac{1}{2\tilde{\mathbf{c}}}\right)\frac{2c\|\widehat{\Gamma}_{\tau 0}^{-1}\|_\infty}{\ell c - 1}\frac{n}{\lambda}\{\widehat{c}_f + \widehat{c}_r\}^2$$

□

Proof of Lemma 4. Note that $\|\widehat{f}\|_\infty^2$ and $\|\widehat{\Gamma}_0^{-1}\|_\infty$ are uniformly bounded with probability going to one. Under the assumption on the design, for $\mathcal{M}$ defined in Lemma 21 we have that $\min_{m\in\mathcal{M}} \phi_{\max}(m \wedge n)$ is uniformly bounded. Thus by Lemma 21

$$\widehat{s}_m \lesssim_P \left[\frac{n\{\widehat{c}_f + \widehat{c}_r\}}{\lambda} + \sqrt{s}\right]^2.$$

The bound then follows from Lemma 18. □

F.1. **Technical Results for Post-Lasso with Estimated Weights.**

Lemma 18 (Performance of the Post-Lasso)**.** *Under Conditions WL, let $\widehat{T}_{\theta\tau}$ denote the support selected by $\widehat{\theta}_\tau$, and $\widetilde{\theta}_\tau$ be the Post-Lasso estimator based on $\widehat{T}_{\theta\tau}$. Then we have for $\widehat{s}_{\theta\tau} = |\widehat{T}_{\theta\tau}|$*

$$\|\widehat{f}_i(m_{\tau i} - x_i'\widetilde{\theta}_\tau)\|_{2,n} \lesssim_P \sqrt{\frac{\phi_{\max}(\widehat{s}_{\theta\tau})}{\phi_{\min}(\widehat{s}_{\theta\tau})}}\frac{c_f}{\min_{i\leqslant n}\widehat{f}_i} + \frac{\sqrt{\widehat{s}_{\theta\tau}}\sqrt{\log p}}{\sqrt{n}\,\phi_{\min}(\widehat{s}_{\theta\tau})\min_{i\leqslant n}\widehat{f}_i} + \min_{\text{support}(\theta)\subseteq\widehat{T}_{\theta\tau}} \|\widehat{f}_i(m_{\tau i} - x_i'\theta)\|_{2,n}$$

Moreover, if in addition λ satisfies (A.37), and $\ell\widehat{\Gamma}_{\tau 0} \leqslant \widehat{\Gamma}_\tau \leqslant u\widehat{\Gamma}_{\tau 0}$ with $u \geqslant 1 \geqslant \ell > 1/c$ in the first stage for Lasso, then we have with high probability

$$\min_{\text{support}(\theta)\subseteq\widehat{T}_{\theta\tau}} \|\widehat{f}_i(m_{\tau i} - x_i'\theta)\|_{2,n} \leqslant 3\{c_f + c_r\} + \left(u + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_{\widehat{\mathbf{c}}}\min_{i\leqslant n}\widehat{f}_i} + 3\bar{f}C\sqrt{s/n}.$$

Proof of Lemma 18. Let $F = \text{diag}(f)$, $\widehat{F} = \text{diag}(\widehat{f})$, $X = [x_1; \ldots; x_n]'$ and for a set of indices $S \subset \{1, \ldots, p\}$ we define $P_S = FX[S](FX[S]'FX[S])^{-1}FX[S]'$ and $\widehat{P}_S = \widehat{F}X[S](X[S]'\widehat{F}'\widehat{F}X[S])^{-1}\widehat{F}X[S]'$ denote the projection matrix on the columns associated with the indices in S. Since $f_i d_i = f_i m_{\tau i} + v_i$ we have that $\widehat{f}_i d_i = \widehat{f}_i m_{\tau i} + v_i\widehat{f}_i/f_i$ and we have

$$\widehat{F}m_\tau - \widehat{F}X\widetilde{\theta}_\tau = (I - \widehat{P}_{\widehat{T}_{\theta\tau}})\widehat{F}m_\tau - \widehat{P}_{\widehat{T}_{\theta\tau}}\widehat{F}F^{-1}v$$

where I is the identity operator. Therefore

$$\|\widehat{F}m_\tau - \widehat{F}X\widetilde{\theta}_\tau\| \leqslant \|(I - \widehat{P}_{\widehat{T}_{\theta\tau}})\widehat{F}m_\tau\| + \|\widehat{P}_{\widehat{T}_{\theta\tau}}\widehat{F}F^{-1}v\|. \tag{F.68}$$

Since $\|\widehat{F}X[\widehat{T}_{\theta\tau}]/\sqrt{n}(X[\widehat{T}_{\theta\tau}]'\widehat{F}'\widehat{F}X[\widehat{T}_{\theta\tau}]/n)^{-1}\| \leqslant \|\widehat{F}^{-1}\|_\infty\sqrt{1/\phi_{\min}(\widehat{s}_{\theta\tau})}$, the last term in (F.68) satisfies

$$\begin{aligned}
\|\widehat{P}_{\widehat{T}_{\theta\tau}}\widehat{F}F^{-1}v\| &\leqslant \|\widehat{F}^{-1}\|_\infty\sqrt{1/\phi_{\min}(\widehat{s}_{\theta\tau})}\ \|X[\widehat{T}_{\theta\tau}]'\widehat{F}^2F^{-1}v/\sqrt{n}\| \\
&\leqslant \|\widehat{F}^{-1}\|_\infty\sqrt{1/\phi_{\min}(\widehat{s}_{\theta\tau})}\left\{\|X[\widehat{T}_{\theta\tau}]'\{\widehat{F}^2 - F^2\}F^{-1}v/\sqrt{n}\| + \|X[\widehat{T}_{\theta\tau}]'Fv/\sqrt{n}\|\right\} \\
&\leqslant \|\widehat{F}^{-1}\|_\infty\sqrt{1/\phi_{\min}(\widehat{s}_{\theta\tau})}\left\{\|X[\widehat{T}_{\theta\tau}]'\{\widehat{F}^2 - F^2\}F^{-1}v/\sqrt{n}\| + \sqrt{\widehat{s}_{\theta\tau}}\|X'Fv/\sqrt{n}\|_\infty\right\}.
\end{aligned}$$

Condition WL(iii) implies that

$$\|X[\widehat{T}_{\theta\tau}]'\{\widehat{F}^2 - F^2\}F^{-1}v/\sqrt{n}\| \leqslant \sup_{\|\alpha\|_0\leqslant\widehat{s}_{\theta\tau},\|\alpha\|\leqslant 1}|\alpha'X[\widehat{T}_{\theta\tau}]'\{\widehat{F}^2 - F^2\}F^{-1}v/\sqrt{n}| \leqslant \sqrt{n}\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\widehat{c}_f.$$

Under Condition WL(iv), by Lemma 12 we have

$$\|X'Fv/\sqrt{n}\|_\infty \lesssim_P \sqrt{\log p}\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[f_i^2x_{ij}^2v_i^2]}.$$

Moreover, Condition WL(iv) also implies $\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[f_i^2x_{ij}^2v_i^2]} \lesssim_P 1$ since $\max_{1\leqslant j\leqslant p}|(\mathbb{E}_n - \bar{\mathrm{E}})[f_i^2x_{ij}^2v_i^2]| \leqslant \delta_n$ with probability $1 - \Delta_n$, and $\max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[f_i^2x_{ij}^2v_i^2] \leqslant \bar{f}^2\bar{c}^2\bar{\mathrm{E}}[x_{ij}^2] \lesssim 1$.

The last statement follows from noting that the Lasso solution provides an upper bound to the approximation of the best model based on $\widehat{T}_{\theta\tau}$, and the application of Lemma 3. □

Lemma 19 (Empirical pre-sparsity for Lasso). *Let $\widehat{T}_{\theta\tau}$ denote the support selected by the Lasso estimator, $\widehat{s}_{\theta\tau} = |\widehat{T}_{\theta\tau}|$, assume $\lambda/n \geqslant c\|\mathbb{E}_n[\widehat{\Gamma}_{\tau 0}^{-1} f_i x_i v_i]\|_\infty$, and $\ell\widehat{\Gamma}_{\tau 0} \leqslant \widehat{\Gamma}_\tau \leqslant u\widehat{\Gamma}_{\tau 0}$ with $u \geqslant 1 \geqslant \ell > 1/c$. Then, for $c_0 = (uc+1)/(\ell c - 1)$ and $\tilde{\mathbf{c}} = (uc+1)/(\ell c - 1)\|\widehat{\Gamma}_{\tau 0}\|_\infty \|\widehat{\Gamma}_{\tau 0}^{-1}\|_\infty$ we have*

$$\sqrt{\widehat{s}_{\theta\tau}} \leqslant 2\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}(1 + 3\|\widehat{f}\|_\infty)\|\widehat{\Gamma}_0^{-1}\|_\infty c_0 \left[\frac{n\{\widehat{c}_f + \widehat{c}_r\}}{\lambda} + \frac{\sqrt{s}\|\widehat{\Gamma}_{\tau 0}\|_\infty}{\kappa_{\tilde{\mathbf{c}}} \min_{i \leqslant n} \widehat{f}_i}\right].$$

Proof of Lemma 19. Let $\widehat{F} = \mathrm{diag}(\widehat{f})$, $R_{\theta\tau} = (r_{\theta\tau 1}, \ldots, r_{\theta\tau n})'$, and $X = [x_1; \ldots; x_n]'$. We have from the optimality conditions that the Lasso estimator $\widehat{\theta}_\tau$ satisfies

$$2\mathbb{E}_n[\widehat{\Gamma}_j^{-1} \widehat{f}_i^2 x_i (d_i - x_i'\widehat{\theta}_\tau)] = \mathrm{sign}(\widehat{\theta}_{\tau j})\lambda/n \quad \text{for each} \quad j \in \widehat{T}_{\theta\tau}.$$

Therefore, noting that $\|\widehat{\Gamma}^{-1}\widehat{\Gamma}^0\|_\infty \leqslant 1/\ell$, we have

$$\begin{aligned}
&\sqrt{\widehat{s}_{\theta\tau}}\lambda = 2\|(\widehat{\Gamma}^{-1}X'\widehat{F}^2(D - X\widehat{\theta}_\tau))_{\widehat{T}_{\theta\tau}}\| \\
&\leqslant 2\|(\widehat{\Gamma}^{-1}X'FV)_{\widehat{T}_{\theta\tau}}\| + 2\|(\widehat{\Gamma}^{-1}X'(\widehat{F}^2 - F^2)F^{-1}V)_{\widehat{T}_{\theta\tau}}\| + 2\|(\widehat{\Gamma}^{-1}X'\widehat{F}^2 R_{\theta\tau})_{\widehat{T}_{\theta\tau}}\| + 2\|(\widehat{\Gamma}^{-1}X'\widehat{F}^2 X(\theta_\tau - \widehat{\theta}_\tau))_{\widehat{T}_{\theta\tau}}\| \\
&\leqslant \sqrt{\widehat{s}_{\theta\tau}}\ \|\widehat{\Gamma}^{-1}\widehat{\Gamma}_0\|_\infty \|\widehat{\Gamma}_{\tau 0}^{-1}X'F'V\|_\infty + 2n\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\|\widehat{\Gamma}^{-1}\|_\infty\{c_f + \|\widehat{F}\|_\infty c_r\} + \\
&2n\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\|\widehat{F}\|_\infty\|\widehat{\Gamma}^{-1}\|_\infty\|\widehat{f}_i x_i'(\widehat{\theta}_\tau - \theta_\tau)\|_{2,n}, \\
&\leqslant \sqrt{\widehat{s}_{\theta\tau}}\ (1/\ell)\ n\|\widehat{\Gamma}_{\tau 0}^{-1}X'FV\|_\infty + 2n\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\frac{\|\widehat{\Gamma}_0^{-1}\|_\infty}{\ell}(c_f + \|\widehat{F}\|_\infty c_r + \|\widehat{F}\|_\infty\|\widehat{f}_i x_i'(\widehat{\theta}_\tau - \theta_\tau)\|_{2,n}),
\end{aligned}$$

where we used that

$$\begin{aligned}
&\|(X'\widehat{F}^2(\theta_\tau - \widehat{\theta}_\tau))_{\widehat{T}_{\theta\tau}}\| \\
&\leqslant \sup_{\|\delta\|_0 \leqslant \widehat{s}_{\theta\tau}, \|\delta\| \leqslant 1} |\delta' X'\widehat{F}^2 X(\theta_\tau - \widehat{\theta}_\tau)| \leqslant \sup_{\|\delta\|_0 \leqslant \widehat{s}_{\theta\tau}, \|\delta\| \leqslant 1} \|\delta' X'\widehat{F}'\|\|\widehat{F}X(\theta_\tau - \widehat{\theta}_\tau)\| \\
&\leqslant \sup_{\|\delta\|_0 \leqslant \widehat{s}_{\theta\tau}, \|\delta\| \leqslant 1} \{\delta' X'\widehat{F}^2 X\delta\}^{1/2}\|\widehat{F}X(\theta_\tau - \widehat{\theta}_\tau)\| \leqslant n\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\|\widehat{f}_i\|_\infty\|\widehat{f}_i x_i'(\theta_\tau - \widehat{\theta}_\tau)\|_{2,n}, \\
&\|(X'(\widehat{F}^2 - F^2)F^{-1}V)_{\widehat{T}_{\theta\tau}}\| \leqslant \sup_{\|\delta\|_0 \leqslant \widehat{s}_{\theta\tau}, \|\delta\| \leqslant 1} |\delta' X'(\widehat{F}^2 - F^2)F^{-1}V| \\
&\leqslant \sup_{\|\delta\|_0 \leqslant \widehat{s}_{\theta\tau}, \|\delta\| \leqslant 1} \|X\delta\|\ \|(\widehat{F}^2 - F^2)F^{-1}V\| \leqslant n\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\widehat{c}_f
\end{aligned}$$

Since $\lambda/c \geqslant \|\widehat{\Gamma}_{\tau 0}^{-1}X'FV\|_\infty$, and by Lemma 3, $\|\widehat{f}_i x_i'(\widehat{\theta}_\tau - \theta_\tau)\|_{2,n} \leqslant 2\{\widehat{c}_f + \widehat{c}_r\} + \left(u + \frac{1}{c}\right)\frac{\lambda\sqrt{s}\|\widehat{\Gamma}_{\tau 0}\|_\infty}{n\kappa_{\tilde{\mathbf{c}}}\min_{i \leqslant n}\widehat{f}_i}$ we have

$$\sqrt{\widehat{s}_{\theta\tau}} \leqslant \frac{2\sqrt{\phi_{\max}(\widehat{s}_{\theta\tau})}\frac{\|\widehat{\Gamma}_0^{-1}\|_\infty}{\ell}\left[\frac{n\widehat{c}_f}{\lambda}(1 + 2\|\widehat{F}\|_\infty) + \frac{n\widehat{c}_r}{\lambda}3\|\widehat{F}\|_\infty + \|\widehat{F}\|_\infty\left(u + \frac{1}{c}\right)\frac{\sqrt{s}\|\widehat{\Gamma}_{\tau 0}\|_\infty}{\kappa_{\tilde{\mathbf{c}}}\min_{i \leqslant n}\widehat{f}_i}\right]}{\left(1 - \frac{1}{c\ell}\right)}.$$

The result follows by noting that $(u + [1/c])/(1 - 1/[\ell c]) = c_0\ell$ by definition of c_0. □

Lemma 20 (Sub-linearity of maximal sparse eigenvalues). *Let M be a semi-definite positive matrix. For any integer $k \geqslant 0$ and constant $\ell \geqslant 1$ we have $\phi_{\max}(\lceil \ell k \rceil)(M) \leqslant \lceil \ell \rceil \phi_{\max}(k)(M)$.*

Lemma 21 (Sparsity bound for Estimated Lasso under data-driven penalty). *Consider the Lasso estimator $\widehat{\theta}_\tau$, let $\widehat{s}_{\theta\tau} = |\widehat{T}_{\theta\tau}|$, and assume that $\lambda/n \geqslant c\|\mathbb{E}_n[\widehat{\Gamma}_{\tau 0}^{-1} f_i x_i v_i]\|_\infty$. Consider the set*

$$\mathcal{M} = \left\{ m \in \mathbb{N} : m > 8\phi_{\max}(m)(1 + 3\|\widehat{f}\|_\infty)^2\|\widehat{\Gamma}_0^{-1}\|_\infty^2 c_0^2 \left[\frac{n\{c_f + c_r\}}{\lambda} + \frac{\sqrt{s}\|\widehat{\Gamma}_{\tau 0}\|_\infty}{\kappa_{\tilde{\mathbf{c}}}\min_{i \leqslant n}\widehat{f}_i}\right]^2 \right\}.$$

Then,

$$\widehat{s}_{\theta\tau} \leqslant 4\left(\min_{m\in\mathcal{M}} \phi_{\max}(m)\right)(1+3\|\widehat{f}\|_\infty)^2\|\widehat{\Gamma}_0^{-1}\|_\infty^2 c_0^2\left[\frac{n\{c_f+c_r\}}{\lambda}+\frac{\sqrt{s}\|\widehat{\Gamma}_{\tau 0}\|_\infty}{\kappa_{\tilde{\mathbf{c}}}\min_{i\leqslant n}\widehat{f}_i}\right]^2.$$

Proof of Lemma 21. Let $L_n = 2(1+3\|\widehat{f}\|_\infty)\|\widehat{\Gamma}_0^{-1}\|_\infty c_0\left[\frac{n\{c_f+c_r\}}{\lambda}+\frac{\sqrt{s}\|\widehat{\Gamma}_{\tau 0}\|_\infty}{\kappa_{\tilde{\mathbf{c}}}\min_{i\leqslant n}\widehat{f}_i}\right]$. Rewriting the conclusion in Lemma 19 we have

$$\widehat{s}_{\theta\tau} \leqslant \phi_{\max}(\widehat{s}_{\theta\tau})L_n^2. \tag{F.69}$$

Consider any $M \in \mathcal{M}$, and suppose $\widehat{s}_{\theta\tau} > M$. Therefore by the sublinearity of the maximum sparse eigenvalue (see Lemma 20)

$$\widehat{s}_{\theta\tau} \leqslant \left\lceil \frac{\widehat{s}_{\theta\tau}}{M}\right\rceil \phi_{\max}(M)L_n^2.$$

Thus, since $\lceil k\rceil \leqslant 2k$ for any $k \geqslant 1$ we have

$$M \leqslant 2\phi_{\max}(M)L_n^2$$

which violates the condition that $M \in \mathcal{M}$. Therefore, we have $\widehat{s}_{\theta\tau} \leqslant M$.

In turn, applying (F.69) once more with $\widehat{s}_{\theta\tau} \leqslant M$ we obtain

$$\widehat{s}_{\theta\tau} \leqslant \phi_{\max}(M)L_n^2.$$

The result follows by minimizing the bound over $M \in \mathcal{M}$. □

Appendix G. Relevant Approximations Rates for $\widehat{f}$

Let $\widehat{Q}(u \mid \tilde{x}) = \tilde{x}'\widehat{\eta}_u$ for $u = \tau - h, \tau + h$. Using a Taylor expansion for the conditional quantile function $Q(\cdot \mid \tilde{x})$, assuming that $\sup_{|\tilde{\tau}-\tau|\leqslant h}|Q'''(\tilde{\tau} \mid \tilde{x})| \leqslant C$ we have

$$|\widehat{Q}'(\tau \mid \tilde{x}) - Q'(\tau \mid \tilde{x})| \leqslant \frac{|Q(\tau+h \mid \tilde{x}) - \tilde{x}'\widehat{\eta}_{\tau+h}| + |Q(\tau-h \mid \tilde{x}) - \tilde{x}'\widehat{\eta}_{\tau-h}|}{h} + Ch^2.$$

In turn, to estimate f_i, the conditional density at $Q(\tau \mid \tilde{x})$, we set $\widehat{f}_i = 1/\widehat{Q}'(\tau \mid \tilde{x}_i)$ which leads to

$$|f_i - \widehat{f}_i| = \frac{|\widehat{Q}'(\tau \mid \tilde{x}_i) - Q'(\tau \mid \tilde{x}_i)|}{\widehat{Q}'(\tau \mid \tilde{x}_i)Q'(\tau \mid \tilde{x}_i)} = (\widehat{f}_i f_i)\cdot|\widehat{Q}'(\tau \mid \tilde{x}_i) - Q'(\tau \mid \tilde{x}_i)|. \tag{G.70}$$

Lemma 22 (Bound Rates for Density Estimator)**.** *Let $\tilde{x} = (d, x)$, suppose that $c \leqslant f_i \leqslant C$, $\sup_\epsilon f'_{\epsilon_i|\tilde{x}_i}(\epsilon \mid \tilde{x}_i) \leqslant \bar{f}' \leqslant C$, $i = 1, \ldots, n$, uniformly in n. Assume further that with probability $1 - \Delta_n$ we have for $u = \tau - h, \tau + h$ that*

$$\|\tilde{x}_i'(\widehat{\eta}_u - \eta_u) + r_{ui}\|_{2,n} \leqslant \frac{C}{\kappa_{\mathbf{c}}}\sqrt{\frac{s\log(p\vee n)}{n}},\ \ \|\widehat{\eta}_u - \eta_u\|_1 \leqslant \frac{C}{\kappa_{\mathbf{c}}^2}\sqrt{\frac{s^2\log(p\vee n)}{n}}\ \ \text{and}\ \ |\widehat{\eta}_{u1} - \eta_{u1}| \leqslant \frac{C}{\kappa_{\mathbf{c}}}\sqrt{\frac{s\log(p\vee n)}{n}}.$$

Then if $\sup_{|\tilde{\tau}-\tau|\leqslant h}|Q'''(\tilde{\tau} \mid \tilde{x})| \leqslant C$, $\max_{i\leqslant n}\|x_i\|_\infty\sqrt{s^2\log(p\vee n)} + \max_{i\leqslant n}|d_i|\sqrt{s\log(p\vee n)} \leqslant \delta_n h\kappa_{\mathbf{c}}^2\sqrt{n}$ and $\max_{u=\tau+h,\tau-h}\|r_{ui}\|_\infty \leqslant h\delta_n$ we have

$$\|f_i - \widehat{f}_i\|_{2,n} \lesssim_P \frac{1}{h\kappa_{\mathbf{c}}}\sqrt{\frac{s\log(n\vee p)}{n}} + h^2, \quad \text{and}$$

$$\max_{i\leqslant n}|f_i-\widehat{f}_i|\lesssim_P \max_{u=\tau+h,\tau-h}\frac{\|r_{ui}\|_\infty}{h}+\frac{\max_{i\leqslant n}\|x_i\|_\infty}{h\kappa_{\mathbf{c}}^2}\sqrt{\frac{s^2\log(n\vee p)}{n}}+\frac{\max_{i\leqslant n}|d_i|_\infty}{h\kappa_{\mathbf{c}}}\sqrt{\frac{s\log(n\vee p)}{n}}+h^2.$$

Proof. Letting $(\delta_\alpha^u;\delta_\beta^u)=\eta_u-\widehat{\eta}_u$ and $\tilde{x}_i=(d_i,x_i')'$ we have that

$$\begin{aligned}|\widehat{f}_i-f_i| &\leqslant |f_i\widehat{f}_i\tfrac{\tilde{x}_i'(\eta_{\tau+h}-\widehat{\eta}_{\tau+h})+r_{g\tau+h,i}-\tilde{x}_i'(\eta_{\tau-h}-\widehat{\eta}_{\tau-h})-r_{g\tau-h,i}}{2h}|+Ch^2\\ &= h^{-1}(f_i\widehat{f}_i)|x_i'\delta_\beta^{\tau+h}+d_i\delta_\alpha^{\tau+h}+r_{g\tau+h,i}-x_i'\delta_\beta^{\tau-h}-d_i\delta_\alpha^{\tau-h}-r_{g\tau-h,i}\}|+Ch^2\\ &\leqslant h^{-1}(f_i\widehat{f}_i)\left\{K_x\|\eta_{\tau+h}\|_1+K_x\|\eta_{\tau-h}\|_1+|d_i|\cdot|\delta_\alpha^{\tau+h}|+|d_i|\cdot|\delta_\alpha^{\tau-h}|+|r_{g\tau+h,i}-r_{g\tau-h,i}|\right\}+Ch^2.\end{aligned}$$

The result follows because for sequences $d_n\to 0, c_n\to 0$ we have $|\widehat{f}_i-f_i|\leqslant|\widehat{f}_if_i|c_n+d_n$ implies that $\widehat{f}_i(1-f_ic_n)\leqslant f_i+d_n$. Since f_i is bounded, $f_ic_n\to 0$ which implies that $\widehat{f}_i$ is bounded. Therefore, $|\widehat{f}_i-f_i|\lesssim c_n+d_n$. We take $d_n=Ch^2\to 0$ and

$$c_n=h^{-1}\left\{K_x\|\eta_{\tau+h}\|_1+K_x\|\eta_{\tau-h}\|_1+|d_i|\cdot|\delta_\alpha^{\tau+h}|+|d_i|\cdot|\delta_\alpha^{\tau-h}|+|r_{g\tau+h,i}-r_{g\tau-h,i}|\right\}\to_P 0$$

by the growth condition.

Moreover, we have

$$\|(\widehat{f}_i-f_i)/f_i\|_{2,n}\lesssim\frac{\|\widehat{f}_i\tilde{x}_i'(\widehat{\eta}_{\tau+h}-\eta_{\tau+h})+\widehat{f}_ir_{g\tau+h,i}\|_{2,n}+\|\widehat{f}_i\tilde{x}_i'(\widehat{\eta}_{\tau-h}-\eta_{\tau-h})+\widehat{f}_ir_{g\tau+h,i}\|_{2,n}}{h}+Ch^2.$$

By the previous result $\widehat{f}_i$ is uniformly bounded from above with high probability. Thus, the result follows by the assumed prediction norm rate $\|\tilde{x}_i'(\widehat{\eta}_u-\eta_u)+r_{ui}\|_{2,n}\lesssim_P(1/\kappa_{\mathbf{c}})\sqrt{s\log(p\vee n)/n}$.

□

Appendix H. Results for Section A.3

Let $(d,z)\in\mathcal{D}\times\mathcal{Z}$. In this section for $\tilde{h}=(\tilde{g},\tilde{\imath})$, where $\tilde{g}$ is a function of variable z, and the instrument $\tilde{\imath}$ is a function on $(d,z)\mapsto\tilde{\imath}(d,z)$ we write

$$\psi_{\tilde{\alpha},\tilde{h}}(y_i,d_i,z_i)=\psi_{\tilde{\alpha},\tilde{g},\tilde{\imath}}(y_i,d_i,z_i)=(\tau-1\{y_i\leqslant\tilde{g}(z_i)+d_i\alpha\})\tilde{\imath}(d_i,x_i)=(\tau-1\{y_i\leqslant\tilde{g}_i+d_i\alpha\})\tilde{\imath}_i.$$

For a fixed $\tilde{\alpha}\in\mathbb{R}$, $\tilde{g}:\mathcal{Z}\to\mathbb{R}$, and $\tilde{\imath}:\mathcal{D}\times\mathcal{Z}\to\mathbb{R}$ we define

$$\Gamma(\tilde{\alpha},\tilde{h}):=\bar{\mathrm{E}}[\psi_{\alpha,h}(y_i,d_i,z_i)]\Big|_{\alpha=\tilde{\alpha},h=\tilde{h}}$$

where the expectation is taken with respect to $\{y_i,i=1,\dots,n\}$ conditionally on $\{d_i,z_i,i=1,\dots,n\}$ is fixed. We use the following notation. Let $\tilde{\imath}_i=\tilde{\imath}(d_i,z_i)$ and $\tilde{g}_i=\tilde{g}(z_i)$, $h_0=(g_\tau,z_0)$ and $\widehat{h}=(\widehat{g},\widehat{z})$. The partial derivative of Γ with respect to α at $(\tilde{\alpha},\tilde{h})$ is denoted by $\Gamma_\alpha(\tilde{\alpha},\tilde{h})$ and the directional derivative with respect to $[\widehat{h}-h_0]$ at $(\tilde{\alpha},\tilde{h})$ is denote as

$$\Gamma_h(\tilde{\alpha},\tilde{h})[\widehat{h}-h_0]=\lim_{t\to 0}\frac{\Gamma(\tilde{\alpha},\tilde{h}+t[\widehat{h}-h_0])-\Gamma(\tilde{\alpha},\tilde{h})}{t}.$$

Proof of Lemma 5. Steps 1-4 we use IQR(i-iii). In Steps 5 and 6 we will also use IQR(iv).

Step 1. (Normality result) We have

$$\begin{aligned}\overbrace{\mathbb{E}_n[\psi_{\check{\alpha}_\tau,\widehat{h}}(y_i,d_i,z_i)]}^{(0)} &= \mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)] + \mathbb{E}_n[\psi_{\check{\alpha}_\tau,\widehat{h}}(y_i,d_i,z_i) - \psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)] \\ &= \underbrace{\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]}_{(I)} + \underbrace{\Gamma(\check{\alpha}_\tau,\widehat{h})}_{(II)} + \underbrace{n^{-1/2}\mathbb{G}_n(\psi_{\check{\alpha}_\tau,\widehat{h}} - \psi_{\check{\alpha}_\tau,h_0})}_{(III)} + \underbrace{n^{-1/2}\mathbb{G}_n(\psi_{\check{\alpha}_\tau,h_0} - \psi_{\alpha_\tau,h_0})}_{(IV)}\end{aligned}$$

Condition IQR(iii), relation (A.43), yields that with probability at least $1-\Delta_n$ we have $|(0)| \lesssim \delta_n n^{-1/2}$.

Step 2 below establishes that $|(II) + \bar{\mathrm{E}}[f_i d_i \iota_{0i}](\check{\alpha}_\tau - \alpha_\tau)| \lesssim_P \delta_n n^{-1/2} + \delta_n|\check{\alpha}_\tau - \alpha_\tau|$.

Condition IQR(iii), relation (A.42), shows that with probability at least $1-\Delta_n$ we have $|(III)| \lesssim \delta_n n^{-1/2}$.

We now proceed to bound term (IV). By Condition IQR(iii) we have with probability at least $1-\Delta_n$ that $|\check{\alpha}_\tau - \alpha_\tau| \leqslant \delta_n$. Observe that

$$\begin{aligned}(\psi_{\alpha,h_0} - \psi_{\alpha_\tau,h_0})(y_i,d_i,z_i) &= (1\{y_i \leqslant g_{\tau i} + d_i\alpha_\tau\} - 1\{y_i \leqslant g_{\tau_i} + d_i\alpha\})\iota_{0i} \\ &= (1\{\epsilon_i \leqslant 0\} - 1\{\epsilon_i \leqslant d_i(\alpha - \alpha_\tau)\})\iota_{0i},\end{aligned}$$

so that $|(\psi_{\alpha,h_0} - \psi_{\alpha_\tau,h_0})(y_i,d_i,z_i)| \leqslant 1\{|\epsilon_i| \leqslant \delta_n|d_i|\}|\iota_{0i}|$ whenever $|\alpha - \alpha_\tau| \leqslant \delta_n$. Since the class of functions $\{(y,d,z) \mapsto (\psi_{\alpha,h_0} - \psi_{\alpha_\tau,h_0})(y,d,z) : |\alpha - \alpha_\tau| \leqslant \delta_n\}$ is a VC subgraph class with VC index bounded by some constant independent of n, using (a version of) Theorem 2.14.1 in [37], we have

$$\sup_{|\alpha-\alpha_\tau|\leqslant\delta_n} |\mathbb{G}_n(\psi_{\alpha,h_0} - \psi_{\alpha_\tau,h_0})| \lesssim_P (\bar{\mathrm{E}}[1\{|\epsilon_i| \leqslant \delta_n|d_i|\}\iota_{0i}^2])^{1/2} \lesssim_P \delta_n^{1/2}.$$

This implies that $|IV| \lesssim_P \delta_n^{1/2} n^{-1/2}$.

Combining the bounds for (0), (II)-(IV) above we have

$$\bar{\mathrm{E}}[f_i d_i \iota_{0i}](\check{\alpha}_\tau - \alpha_\tau) = \mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)] + O_P(\delta_n^{1/2} n^{-1/2}) + O_P(\delta_n)|\check{\alpha}_\tau - \alpha_\tau|. \tag{H.71}$$

Note that $\mathbb{U}_n(\tau) = \{\bar{\mathrm{E}}[\psi^2_{\alpha_\tau,h_0}(y_i,d_i,z_i)]\}^{-1/2}\sqrt{n}\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]$ and $\bar{\mathrm{E}}[\psi^2_{\alpha_\tau,h_0}(y_i,d_i,z_i)] = \tau(1-\tau)\bar{\mathrm{E}}[\iota_{0i}^2]$ so that the first representation result follows from (H.71). Since $\bar{\mathrm{E}}[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)] = 0$ and $\bar{\mathrm{E}}[\iota_{0i}^2] \leqslant C$, by the Lyapunov CLT we have

$$\sqrt{n}(I) = \sqrt{n}\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)] \rightsquigarrow N(0, \bar{\mathrm{E}}[\tau(1-\tau)\iota_{0i}^2])$$

and $\mathbb{U}_n(\tau) \rightsquigarrow N(0,1)$ follows by noting that $|\bar{\mathrm{E}}[f_i d_i \iota_{0i}]| \geqslant c > 0$.

Step 2. (Bounding $\Gamma(\alpha,\widehat{h})$ for $|\alpha - \alpha_\tau| \leqslant \delta_n$ which covers (II)) We have

$$\begin{aligned}\Gamma(\alpha,\widehat{h}) &= \Gamma(\alpha,h_0) + \Gamma(\alpha,\widehat{h}) - \Gamma(\alpha,h_0) \\ &= \Gamma(\alpha,h_0) + \{\Gamma(\alpha,\widehat{h}) - \Gamma(\alpha,h_0) - \Gamma_h(\alpha,h_0)[\widehat{h} - h_0]\} + \Gamma_h(\alpha,h_0)[\widehat{h} - h_0].\end{aligned} \tag{H.72}$$

Because $\Gamma(\alpha_\tau,h_0) = 0$, by Taylor expansion there is some $\tilde{\alpha} \in [\alpha_\tau,\alpha]$ such that

$$\Gamma(\alpha,h_0) = \Gamma(\alpha_\tau,h_0) + \Gamma_\alpha(\tilde{\alpha},h_0)(\alpha - \alpha_\tau) = \{\Gamma_\alpha(\alpha_\tau,h_0) + \eta_n\}(\alpha - \alpha_\tau)$$

where $|\eta_n| \leqslant \delta_n\mathbb{E}_n[|d_i^2\iota_{0i}|] \lesssim_P \delta_n C$ by relation (H.79) in Step 4 and moment conditions in IQR(i).

Combining the argument above with relations (H.74), (H.75) and (H.77) in Step 3 below we have

$$\begin{aligned}\Gamma(\alpha,\widehat{h}) &= \Gamma_h(\alpha_\tau,h_0)[\widehat{h}-h_0]+\Gamma(\alpha_\tau,h_0)+\{\Gamma_\alpha(\alpha_\tau,h_0)+O_P(\delta_n\bar{\mathrm{E}}[|d_i^2\iota_{0i}|])\}(\alpha-\alpha_\tau)+O_P(\delta_n n^{-1/2})\\ &= \Gamma_\alpha(\alpha_\tau,h_0)(\alpha-\alpha_\tau)+O_P(\delta_n|\alpha-\alpha_\tau|\bar{\mathrm{E}}[|d_i^2\iota_{0i}|]+\delta_n n^{-1/2})\end{aligned} \tag{H.73}$$

Step 3. (Relations for Γ_h) The directional derivative Γ_h with respect the direction $\widehat{h}-h_0$ at a point $\tilde{h}=(\tilde{g},\tilde{z})$ is given by

$$\Gamma_h(\alpha,\tilde{h})[\widehat{h}-h_0] \quad = -\mathbb{E}_n[f_{\epsilon_i|d_i,z_i}(d_i(\alpha-\alpha_\tau)+\tilde{g}_i-g_{\tau i})\tilde{\iota}_{0i}\{\widehat{g}_i-g_{\tau i}\}]+\bar{\mathrm{E}}[(\tau-1\{y_i\leqslant\tilde{g}_i+d_i\alpha\})\{\widehat{\iota}_i-\iota_{0i}\}]$$

Note that when Γ_h is evaluated at (α_τ,h_0) we have with probability $1-\Delta_n$

$$|\Gamma_h(\alpha_\tau,h_0)[\widehat{h}-h_0]| = |-\mathbb{E}_n[f_i\iota_{0i}\{\widehat{g}_i-g_{\tau i}\}]| \leqslant \delta_n n^{-1/2} \tag{H.74}$$

by Condition IQR(iii) (A.41) and by $P(y_i\leqslant g_{\tau i}+d_i\alpha_\tau \mid d_i,z_i)=\tau$. The expression for Γ_h also leads to the following bound

$$\begin{aligned}&\left|\Gamma_h(\alpha,h_0)[\widehat{h}-h_0]-\Gamma_h(\alpha_\tau,h_0)[\widehat{h}-h_0]\right| =\\ &= |\mathbb{E}_n[\{f_{\epsilon_i|d_i,z_i}(0)-f_{\epsilon_i|d_i,z_i}(d_i(\alpha-\alpha_\tau))\}\iota_{0i}\{\widehat{g}_i-g_{\tau i}\}]+\mathbb{E}_n[\{F_i(0)-F_i(d_i(\alpha-\alpha_\tau))\}\{\widehat{\iota}_i-\iota_{0i}\}]|\\ &\leqslant \mathbb{E}_n[|\alpha-\alpha_\tau|\ \bar{f}'|d_i\iota_{0i}|\ |\widehat{g}_i-g_{\tau i}|]+\mathbb{E}_n[\bar{f}|(\alpha-\alpha_\tau)d_i|\ |\widehat{\iota}_i-\iota_{0i}|]\\ &\leqslant |\alpha-\alpha_\tau|\cdot\|\widehat{g}_i-g_{\tau i}\|_{2,n}\{\bar{f}'\mathbb{E}_n[\iota_{0i}^2d_i^2]\}^{1/2}+\bar{f}|\alpha-\alpha_\tau|\cdot\{\mathbb{E}_n[(\widehat{\iota}_i-\iota_{0i})^2]\}^{1/2}\{\mathbb{E}_n[d_i^2]\}^{1/2}\\ &\lesssim_P |\alpha-\alpha_\tau|\delta_n\end{aligned} \tag{H.75}$$

The second directional derivative Γ_{hh} at $\tilde{h}=(\tilde{g},\tilde{\iota})$ with respect to the direction $\widehat{h}-h_0$ can be bounded by

$$\begin{aligned}\left|\Gamma_{hh}(\alpha,\tilde{h})[\widehat{h}-h_0,\widehat{h}-h_0]\right| &= \Big|-\mathbb{E}_n[f'_{\epsilon_i|d_i,z_i}(d_i(\alpha-\alpha_\tau)+\tilde{g}_i-g_{\tau i})\tilde{\iota}_i\{\widehat{g}_i-g_{\tau i}\}^2]\\ &\quad +2\mathbb{E}_n[f_{\epsilon_i|d_i,z_i}(d_i(\alpha-\alpha_\tau)+\tilde{g}_i-g_{\tau i})\{\widehat{g}_i-g_{\tau i}\}\{\widehat{\iota}_i-\iota_{0i}\}]\Big|\\ &\leqslant \bar{f}'\max_{i\leqslant n}|\tilde{\iota}_i|\|\widehat{g}_i-g_{\tau i}\|_{2,n}^2+2\bar{f}\|\widehat{g}_i-g_{\tau i}\|_{2,n}\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}.\end{aligned} \tag{H.76}$$

In turn, since $\tilde{h}\in[h_0,\widehat{h}]$, $|\tilde{\iota}(d_i,z_i)|\leqslant|\iota_0(d_i,z_i)|+|\widehat{\iota}(d_i,z_i)-\iota_0(d_i,z_i)|$, we have that

$$\begin{aligned}\left|\Gamma(\alpha,\widehat{h})-\Gamma(\alpha,h_0)\right. &\left.-\Gamma_h(\alpha,h_0)\left[\widehat{h}-h_0\right]\right| \leqslant \sup_{\tilde{h}\in[h_0,\widehat{h}]}\left|\Gamma_{hh}(\alpha,\tilde{h})\left[\widehat{h}-h_0,\widehat{h}-h_0\right]\right|\\ &\leqslant \bar{f}'\left(\max_{i\leqslant n}\{|\iota_{0i}|+|\widehat{\iota}_i-\iota_{0i}|\}\right)\|\widehat{g}_i-g_{\tau i}\|_{2,n}^2+\\ &+2\bar{f}\|\widehat{g}_i-g_{\tau i}\|_{2,n}\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}\\ &\lesssim_P \delta_n n^{-1/2}\end{aligned} \tag{H.77}$$

where the last relation is assumed in Condition IQR(iii).

Step 4. (Relations for Γ_α) By definition of Γ, its derivative with respect to α at $(\alpha,\tilde{h})$ is

$$\Gamma_\alpha(\alpha,\tilde{h}) = -\mathbb{E}_n[f_{\epsilon_i|d_i,z_i}(d_i(\alpha-\alpha_\tau)+\tilde{g}_i-g_{\tau i})d_i\tilde{\iota}_i].$$

Therefore, when the function above is evaluated at $\alpha=\alpha_\tau$ and $\tilde{h}=h_0$, since for $f_{\epsilon_i|d_i,z_i}(0)=f_i$ we have

$$\Gamma_\alpha(\alpha_\tau,h_0) = -\mathbb{E}_n[f_id_i\iota_{0i}] = -\bar{\mathrm{E}}[f_id_i\iota_{0i}]-(\mathbb{E}_n-\bar{\mathrm{E}})[f_id_i\iota_{0i}] = -\bar{\mathrm{E}}[f_id_i\iota_{0i}]+O_P(n^{-1/2}). \tag{H.78}$$

Moreover, Γ_α also satisfies

$$\begin{aligned}|\Gamma_\alpha(\alpha,h_0)-\Gamma_\alpha(\alpha_\tau,h_0)| &= \left|\mathbb{E}_n[f_{\epsilon_i|d_i,z_i}(d_i(\alpha-\alpha_\tau))\iota_{0i}d_i]-\mathbb{E}_n[f_{\epsilon_i|d_i,z_i}(0)\iota_{0i}d_i]\right| \\ &\leqslant |\alpha-\alpha_\tau|\bar{f}'\mathbb{E}_n[|d_i^2\iota_{0i}|]=|\alpha-\alpha_\tau|\bar{f}'O_P(\{\bar{\mathrm{E}}[d_i^4]\}^{1/2}\{\bar{\mathrm{E}}[\iota_{0i}^2]\}^{1/2})\end{aligned} \tag{H.79}$$

since $\max_{i\leqslant n}\mathrm{E}[d_i^4]\vee\mathrm{E}[d_i^4]\leqslant C$ by IQR(i).

Step 5. (Estimation of Variance) First note that

$$\begin{aligned}&|\mathbb{E}_n[\widehat{f}_i d_i\widehat{\iota}_i]-\bar{\mathrm{E}}[f_i d_i\iota_{0i}]| \\ &=|\mathbb{E}_n[\widehat{f}_i d_i\widehat{\iota}_i]-\mathbb{E}_n[f_i d_i\iota_{0i}]|+|\mathbb{E}_n[f_i d_i\iota_{0i}]-\bar{\mathrm{E}}[f_i d_i\iota_{0i}]| \\ &\leqslant |\mathbb{E}_n[(\widehat{f}_i-f_i)d_i\widehat{\iota}_i]|+|\mathbb{E}_n[f_i d_i(\widehat{\iota}_i-\iota_{0i})]|+|\mathbb{E}_n[f_i d_i\iota_{0i}]-\bar{\mathrm{E}}[f_i d_i\iota_{0i}]| \\ &\leqslant |\mathbb{E}_n[(\widehat{f}_i-f_i)d_i(\widehat{\iota}_i-\iota_{0i})]|+|\mathbb{E}_n[(\widehat{f}_i-f_i)d_i\iota_{0i}]| \\ &+\|f_i d_i\|_{2,n}\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}+|\ \mathbb{E}_n[f_i d_i\iota_{0i}]-\bar{\mathrm{E}}[f_i d_i\iota_{0i}]| \\ &\lesssim_P \|(\widehat{f}_i-f_i)d_i\|_{2,n}\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}+\|\widehat{f}_i-f_i\|_{2,n}\|d_i\iota_{0i}\|_{2,n} \\ &+\|f_i d_i\|_{2,n}\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}+|\mathbb{E}_n[f_i d_i\iota_{0i}]-\bar{\mathrm{E}}[f_i d_i\iota_{0i}]| \\ &\lesssim_P \delta_n\end{aligned} \tag{H.80}$$

because $f_i,\widehat{f}_i\leqslant C$, $\bar{\mathrm{E}}[d_i^4]\leqslant C$, $\bar{\mathrm{E}}[\iota_{0i}^4]\leqslant C$ by Condition IQR(ii) and Conditions IQR(iii) and (iv).

Next we proceed to control the other term of the variance. We have

$$\begin{aligned}&|\ \|\psi_{\check{\alpha}_\tau,\widehat{h}}(y_i,d_i,z_i)\|_{2,n}-\|\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)\|_{2,n}|\leqslant\|\psi_{\check{\alpha}_\tau,\widehat{h}}(y_i,d_i,z_i)-\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)\|_{2,n} \\ &\leqslant\|\psi_{\check{\alpha}_\tau,\widehat{h}}(y_i,d_i,z_i)-(\tau-1\{y_i\leqslant d_i\check{\alpha}_\tau+\tilde{g}_i\})\iota_{0i}\|_{2,n}+\|(\tau-1\{y_i\leqslant d_i\check{\alpha}_\tau+\tilde{g}_i\})\iota_{0i}-\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)\|_{2,n} \\ &\leqslant\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}+\|(1\{y_i\leqslant d_i\alpha_\tau+g_{\tau i}\}-1\{y_i\leqslant d_i\check{\alpha}_\tau+\tilde{g}_i\})\iota_{0i}\|_{2,n} \\ &\leqslant\|\widehat{\iota}_i-\iota_{0i}\|_{2,n}+\|\iota_{0i}^2\|_{2,n}^{1/2}\|1\{|\epsilon_i|\leqslant|d_i(\alpha_\tau-\check{\alpha}_\tau)+g_{\tau i}-\tilde{g}_i|\}\|_{2,n}^{1/2} \\ &\lesssim_P\delta_n\end{aligned} \tag{H.81}$$

by IQR(ii) and IQR(iv). Also, $|\mathbb{E}_n[\psi^2_{\alpha_\tau,h_0}(y_i,d_i,z_i)]-\bar{\mathrm{E}}[\psi^2_{\alpha_\tau,h_0}(y_i,d_i,z_i)]|\lesssim_P\delta_n$ by independence and bounded moment conditions in Condition IQR(ii).

Step 6. (Main Step for χ^2) Note that the denominator of $L_n(\alpha_\tau)$ was analyzed in relation (H.81) of Step 5. Next consider the numerator of $L_n(\alpha_\tau)$. Since $\Gamma(\alpha_\tau,h_0)=\bar{\mathrm{E}}[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]=0$ we have

$$\mathbb{E}_n[\psi_{\alpha_\tau,\widehat{h}}(y_i,d_i,z_i)]=(\mathbb{E}_n-\bar{\mathrm{E}})[\psi_{\alpha_\tau,\widehat{h}}(y_i,d_i,z_i)-\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]+\Gamma(\alpha_\tau,\widehat{h})+\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)].$$

By Condition IQR(iii) and (H.73) with $\alpha=\alpha_\tau$, it follows that

$$|(\mathbb{E}_n-\bar{\mathrm{E}})[\psi_{\alpha_\tau,\widehat{h}}(y_i,d_i,z_i)-\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]|\leqslant\delta_n n^{-1/2}\ \text{ and }\ |\Gamma(\alpha_\tau,\widehat{h})|\lesssim_P\delta_n n^{-1/2}.$$

The identity $nA_n^2=nB_n^2+n(A_n-B_n)^2+2nB_n(A_n-B_n)$ for $A_n=\mathbb{E}_n[\psi_{\alpha_\tau,\widehat{h}}(y_i,d_i,x_i)]$ and $B_n=\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,x_i)]\lesssim_P\{\tau(1-\tau)\bar{\mathrm{E}}[\iota_{0i}^2]\}^{1/2}n^{-1/2}$ yields

$$\begin{aligned}nL_n(\alpha_\tau) &= \frac{n|\mathbb{E}_n[\psi_{\alpha_\tau,\widehat{h}}(y_i,d_i,z_i)]|^2}{\mathbb{E}_n[\psi^2_{\alpha_\tau,\widehat{h}}(y_i,d_i,z_i)]} \\ &= \frac{n|\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]|^2+O_P(\delta_n)}{\bar{\mathrm{E}}[\tau(1-\tau)\iota_{0i}^2]+O_P(\delta_n)}=\frac{n|\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i,d_i,z_i)]|^2}{\bar{\mathrm{E}}[\tau(1-\tau)\iota_{0i}^2]}+O_P(\delta_n)\end{aligned}$$

since $\tau(1-\tau)\bar{\mathrm{E}}[\iota_{0i}^2]$ is bounded away from zero because $\underline{C} \leqslant |\bar{\mathrm{E}}[f_i d_i \iota_{0i}]| = |\bar{\mathrm{E}}[v_i \iota_{0i}]| \leqslant \{\bar{\mathrm{E}}[v_i^2]\bar{\mathrm{E}}[\iota_{0i}^2]\}^{1/2}$ and $\bar{\mathrm{E}}[v_i^2]$ is bounded above uniformly. The result then follows since $\sqrt{n}\mathbb{E}_n[\psi_{\alpha_\tau,h_0}(y_i, d_i, z_i)] \rightsquigarrow N(0, \tau(1-\tau)\bar{\mathrm{E}}[\iota_{0i}^2])$.

□