Online Appendix: Inference for Support Vector Regression under ℓ_1 Regularization

Yuehao Bai,*Hung Ho,†Guillaume A. Pouliot,‡Joshua Shea§

A Additional results

A.1 Constructing the ℓ_1-SVR regression rankscore test statistic

We collect a characterization of the complementary slackness condition for the ℓ_1-SVR problem which will be key in the design of our proposed test statistic.

Lemma A.1. *The solution to* (6) *of the main paper satisfies:*

(a) *If* $Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n > \epsilon$, *then* $\hat{a}_i^- = -1$ *and* $\hat{a}_i^+ = 1$.

(b) *If* $Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n < -\epsilon$, *then* $\hat{a}_i^- = -1$ *and* $\hat{a}_i^+ = -1$.

(c) *If* $0 < |Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n| < \epsilon$, *then* $\hat{a}_i^+ = \hat{a}_i^- = 0$.

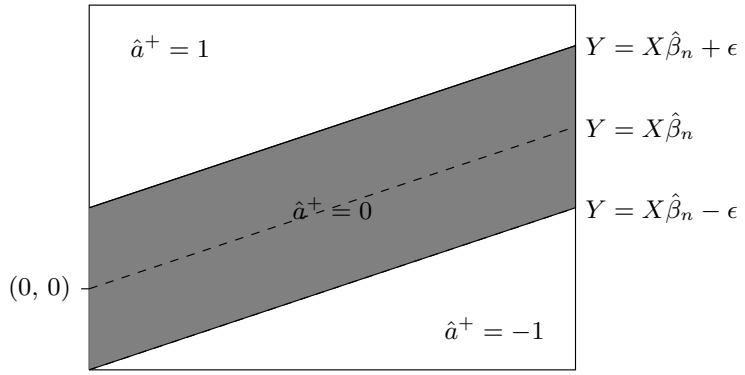

Figure 1: $\hat{a}^+$ in Different Regions of Regression Residuals $Y - X\hat{\beta}_n$

Lemma A.1 provides an alternative intuitive interpretation for $T_n(\mathbf{W}_n, \gamma_0)$ in (7) of the main paper. Since $\lambda_n \to 0$, we intuitively view it as 0. First, as shown in Lemma A.1 and Figure 1, $\hat{a}_i^+$ is a monotonic transformation of the residuals $\hat{u}_i = Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n$. If H_0 holds, i.e., $\gamma(P) = \gamma_0$, then we expect a low correlation between Z and $\hat{u}$, and any monotonic transformation of $\hat{u}$. Therefore $\mathbf{Z}_n'\hat{a}^+$ should be small under the null, but larger the more $\gamma(P)$ differs from γ_0.

*Department of Economics, University of Michigan, 611 Tappan Avenue, Lorch Hall, Ann Arbor, MI 48109, yuehaob@umich.edu.

†The Wharton School, University of Pennsylvania, 3733 Spruce Street, Philadelphia, PA 19104, hqdh@wharton.upenn.edu.

‡Harris School of Public Policy, University of Chicago, 1307 East 60th Street, Chicago, IL 60637, guillaumepouliot@uchicago.edu.

§Department of Economics, University of Chicago, 1126 East 59th Street, Saieh Hall, Chicago IL 60637, jkcshea@uchicago.edu.

Remark A.1. *The complementary slackness conditions in* (5) *and* (6) *are the key ingredients in the construction of the test statistic in* (7)*. These conditions are summarized in Lemma A.1. Figure 1 displays the regions defined by regression residuals and the corresponding values of* $\hat{a}^+$*. For simplicity, we assume* $d_x = 2$*, the intercept is 0, and* $\gamma_0 = 0$*. Note that unit* i *contributes to* $T_n(\mathbf{W}_n, \gamma_0)$ *only when* $|Y_i - X_i\hat{\beta}_n| > \epsilon$*. The graph is different from that under the median regression, where the shaded region in which* $\hat{a}_i^+ = 0$ *collapses to a single line.* ■

Remark A.2. *The* ℓ_1*-SVR regression rankscore test is equivalent to the median regression rankscore test when* $\epsilon = \lambda = 0$ *and* $\hat{p}_n$ *is set to* $\frac{1}{2}$*. See Figure 2.* ■

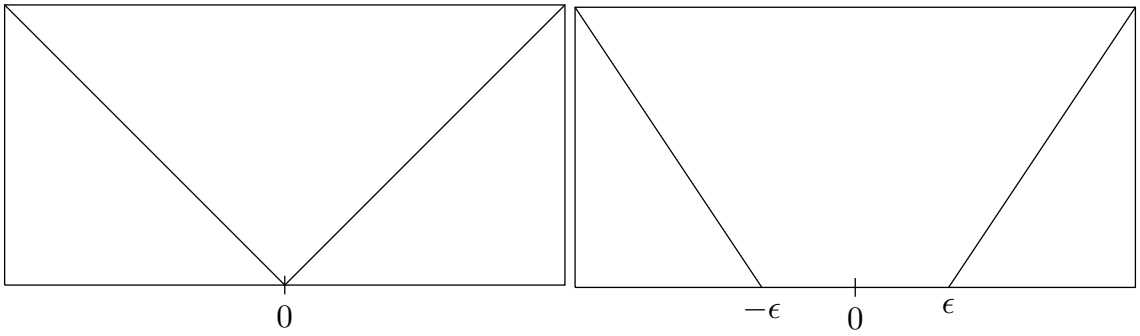

Figure 2: Loss Functions for Median Regression (QR) and for Support Vector Regression (SVR)

Remark A.3. *Let* $\theta = (\beta, \gamma)$ *and* $V = (X, Z)$*. Then*

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N\Big(0, \frac{1}{2}\, E[V_i V_i' f_Y(V_i'\theta - \epsilon \mid V_i)]^{-1} \times E[V_i V_i' F_Y(V_i'\theta - \epsilon \mid V_i)] \times E[V_i V_i' f_Y(V_i'\theta - \epsilon \mid V_i)]^{-1}\Big),$$

from which the Wald confidence interval obtains. See the supplementary appendix for the derivation. However, as shown in Figure 1 of the main paper, Wald confidence intervals are highly sensitive to the bandwidth used for density estimation. ■

Remark A.4. *If* P *violates Assumption I.2(b) but satisfies Assumption I.2(a), it is straightforward to construct a consistent estimator of the asymptotic variance in* (11) *by using the law of iterated expectations. This estimator involves conditional distributions rather than densities, thus does not require a choice of bandwidth. The statistic studentized by the estimate of the asymptotic variance will be asymptotically exact. If* P *violates Assumption I.2(a), it is also possible to construct consistent estimators of the asymptotic variance in* (11)*. The statistic studentized by any consistent estimator of the asymptotic variance will also be asymptotically exact. However, the studentization involves density estimation, which is why Assumption I.2(a) is important for our purposes. See, for example, Powell (1991).* ■

Monotonicity of the test statistic is essential to the tractability of the inversion procedure, as it limits the procedure to a search for the two points where the test statistic, as a function of the posited null parameter, crosses the critical value. Note that without monotonicity, we cannot construct the confidence region as an interval. Error bars are obtained by inverting the test for each individual covariate. As discussed above, these are well-defined only insofar as the individual tests invert to produce intervals—as opposed to disconnected regions. Such a guarantee is given in Theorem A.1.

Theorem A.1. *If* $d_z = 1$*, then* $T_n(\mathbf{W}_n, \gamma_0)$ *in* (7) *is monotonically decreasing in* γ_0*.*

A.2 Additional simulations

For each distribution of errors in the simulation presented in the main paper, we also consider a model that restricts the support of the error terms so that

$$\tau \sim \text{Unif}([0, 0.4] \cup [0.6, 1]).$$

The restricted support model may be thought of as an extreme version of the data generating processes which SVR is meant for when conceived as a regression extension of SVM classification. In addition, we investigate the robustness of the homoskedastic SVR regression rankscore test to heteroskedastic errors.

Table 1 presents the rejection probabilities for the full set of simulations. Columns 1–4 present the simulations in which the errors are homoskedastic. In all the cases of unrestricted support for the errors, the distributions are centered and scaled to be mean zero with a standard deviation of 15.[1] Upon restricting the support, the errors are normalized to be mean zero, but their standard deviation may change. Columns 1 and 3 indicate that the size properties of the SVR and median regression rankscore tests are about equal under homoskedasticity. However, columns 2 and 4 suggest that the SVR regression rankscore test has better power properties than median regression, the former outperforming the latter in 7 of the 10 settings.

Columns 5–8 of Table 1 present the simulations in which the errors are heteroskedastic, their variance being determined by the covariate X.[2] To account for the heteroskedastic errors, we obtain a consistent estimate of the variance expression in (11) using the methodology of Powell (1991). This requires density estimation of the errors, for which the tuning parameters are shown in Table 2.[3] The test statistic is then studentized using the variance estimate. We keep ϵ the same as in the homoskedastic simulations. As before, columns 5–8 of Table 1 suggest that the size properties of the two tests are roughly equal under heteroskedasticity, whereas the SVR regression rankscore test exhibits greater power in 9 of the 10 cases.

To gauge the robustness of the SVR regression rankscore test, columns 9–12 present simulations in which the homoskedastic test statistic is used with heteroskedastic errors. Similar to the earlier results, columns 9 and 11 suggest that both tests have similar size properties. However, columns 10 and 12 reveal that the SVR regression rankscore test demonstrates greater statistical power, having higher rejection rates in 8 of the 10 simulations. Similar robust behavior in heteroskedastic environments of regression rankscore tests constructed under homoskedasticity assumptions are documented using real and simulated data in Bai, Pouliot and Shaikh (2019) and Pouliot (2020).

In each iteration of the simulations in Table 1, we can construct the confidence interval by inverting (10). The results in Table 1 suggest the SVR regression rankscore test has greater power against the alternatives considered when inverting the test, as compared to the median regression rankscore test. From the duality between hypothesis testing and inference, this suggests tighter confidence intervals under the former test procedure. Indeed, we find this to be the case. Figures 3–5 present the average confidence interval for each simulation where $\gamma = 0.5$. In all three figures, the results closely align with those of Table 1. That is, for the data generating process where the SVR regression rankscore test has greater power than the median regression rankscore test, the confidence interval of the former is narrower than that of the latter. The reduction in the error bars becomes rather substantial when we restrict the support of the error terms. This is to be expected, as the SVR loss function is able to account for the restricted support of the error term, whereas the median regression loss function cannot (see Figure 2).

It is rather remarkable that modifications to the quantile regression procedure intended for robustness deliver greater inference accuracy. This naturally suggests using the SVR regression rankscore test for

[1] The mixture consists of evenly weighted Gaussian distributions centered around 10 and -10, with the same scale parameter. Errors under the Student's t distribution are drawn from the counterpart with 10 degrees of freedom, and then rescaled. Errors under the χ^2 distribution are drawn from the counterpart with 3 degrees of freedom, and then rescaled and recentered.

[2] The scale parameters of the distributions are set equal to a normalized value of X. For the Gaussian, Laplace, and mixture distributions, X is normalized to have a standard deviation of 1 and is then recentered so its mean is equal to the scale parameter required for the error distribution to have a standard deviation of 15. For the Student's t and χ^2 distributions, X is instead recentered around the degrees of freedom stated in Footnote 1. In the rare event that the normalized X falls below 0, its absolute value is taken. While the distributions of parameters determining the standard deviation of the errors are centered around the value that would correspond to a standard deviation of 15, the standard deviations of the actual error terms need not be 15.

[3] The bandwidth used in density estimation is $\kappa\left(\Phi^{-1}(0.5 + hn^{-\frac{1}{3}}) - \Phi^{-1}(0.5 - hn^{-\frac{1}{3}})\right)$, where h and κ are constants, and Φ is the standard normal CDF.

Table 1: Rejection Probabilities for Different Distributions of the Errors, Under Homo/Heteroskedasticity

	Homoskedastic				Heteroskedastic				Homo. stat, hetero. data			
	SVR		QR		SVR		QR		SVR		QR	
Distribution	$\gamma=0$	$\gamma=0.5$	$\gamma=0$	$\gamma=0.5$	$\gamma=0$	$\gamma=0.5$	$\gamma=0$	$\gamma=0.5$	$\gamma=0$	$\gamma=0.5$	$\gamma=0$	$\gamma=0.5$
Gaussian	5.7	34.6	5.4	31.3	4.5	34.9	4.7	30.3	5.5	37.0	5.5	32.4
Gaussian, restricted	5.5	15.7	5.7	13.1	4.7	15.5	4.1	10.9	5.5	17.8	5.4	12.6
Laplace	5.7	64.8	5.3	65.4	4.2	63.6	4.7	67.4	5.5	67.9	5.7	70.4
Laplace, restricted	5.7	23.5	5.6	17.8	4.2	25.3	3.9	17.1	5.3	30.1	5.1	20.3
Mixture	4.6	37.9	4.3	30.6	4.5	37.7	4.4	32.1	4.7	38.9	4.7	32.7
Mixture, restricted	5.2	16.5	5.0	14.5	4.8	18.5	4.6	11.7	5.1	19.1	4.8	13.5
Student's t	5.7	40.1	5.4	35.9	4.4	37.0	4.6	31.7	5.7	40.0	5.3	35.9
Student's t, restricted	5.6	17.2	5.7	14.0	4.8	14.1	4.7	11.7	5.4	16.9	5.6	14.0
χ^2	5.1	39.3	5.4	39.6	4.7	16.4	3.7	15.4	4.7	16.9	4.4	17.3
χ^2, restricted	5.4	14.2	5.4	14.8	4.8	16.1	4.0	15.6	5.5	16.6	4.7	16.4

Table 2: Tuning Parameters for Heteroskedastic Test Statistic

	Unrestricted		Restricted	
Distribution	h	κ	h	κ
Gaussian	1	1.75	2.5	2
Laplace	1	0.75	1.5	1.75
Mixture	2	2	2.75	2.5
Student's t	1	1	2	2.75
χ^2	1.5	1.5	2.75	1.5

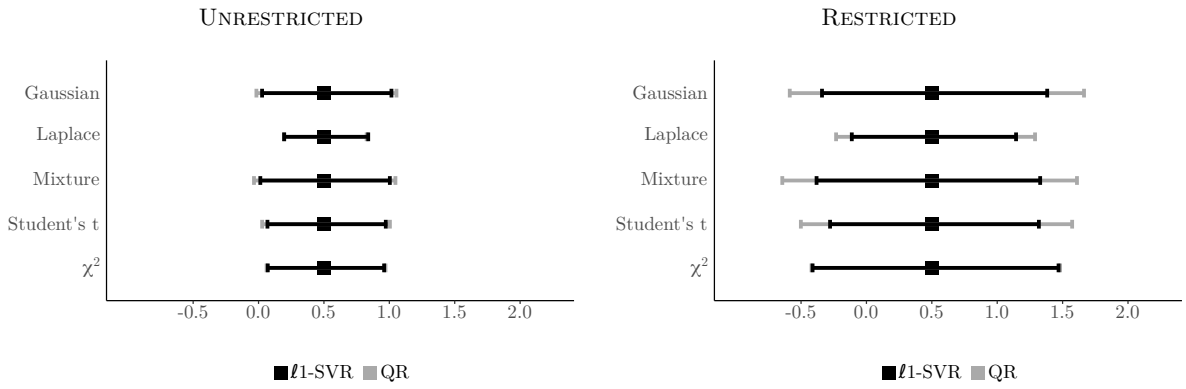

Figure 3: Confidence Intervals Under Homoskedasticity

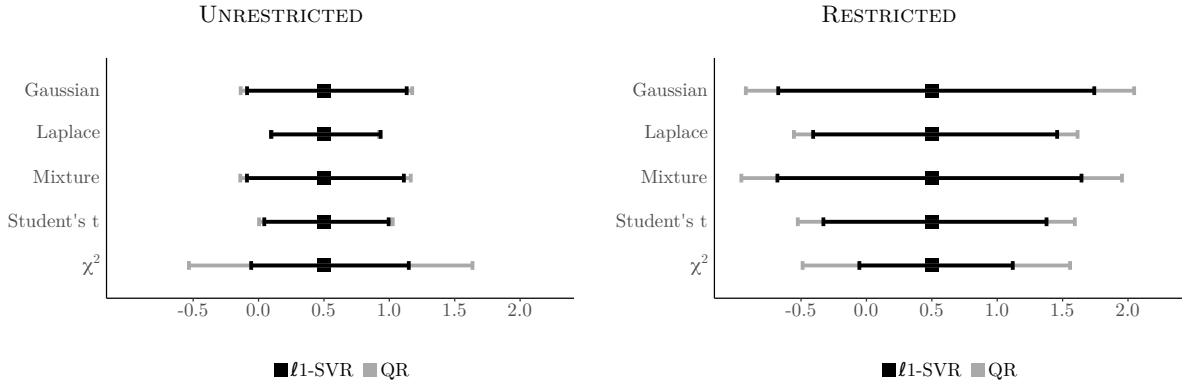

Figure 4: Confidence Intervals Under Heteroskedasticity

inference in standard quantile regression analysis, even if the point estimate is obtained using quantile regression.

A.3 Additional comparisons between ℓ_1-SVR regression rankscore confidence intervals and Wald confidence intervals

Figure 6 compares the coverage probabilities of the 95% confidence intervals constructed from the SVR regression rankscore test and the Wald test. The comparisons are made for each family of error distributions considered in the simulation, with the errors being heteroskedastic and having unrestricted support. A range of bandwidths for estimating the density of the errors is considered. Across all the error distributions, the SVR regression rankscore confidence intervals have the correct coverage probability and are robust to the choice of bandwidth. In contrast, the Wald confidence intervals have coverage probabilities that fall well below the nominal level and are sensitive to the choice of bandwidth.

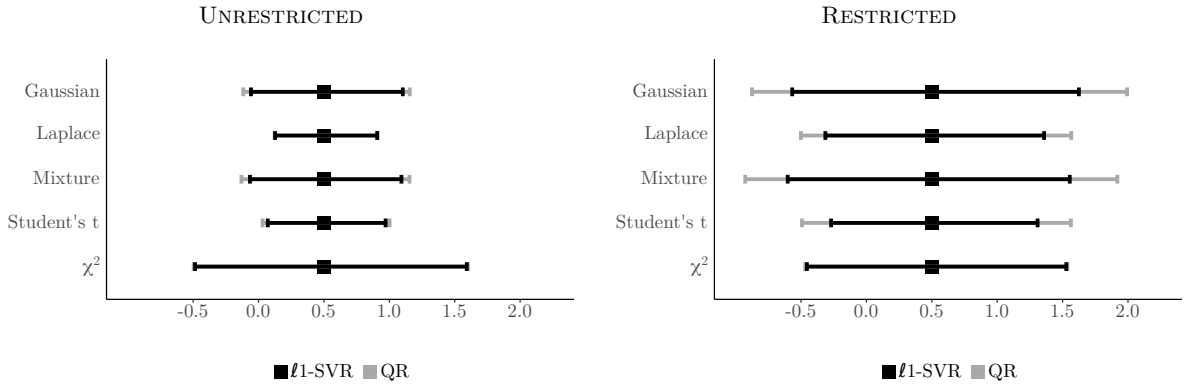

Figure 5: Confidence Intervals Under Heteroskedasticity Using Homoskedastic Test Statistic

B Proofs and derivations

Since all results are derived under the null that $\gamma(P) = \gamma_0$, we assume without loss of generality that $\gamma(P) = \gamma_0 = 0$. We use $a \lesssim b$ to denote that there exists $l > 0$ such that $a \leq lb$. We use $||$ to denote the Euclidean norm.

B.1 Derivation of Equation (3)

Equation (3) can be derived from equation (2) based on the following observation. For any $x \in \mathbb{R}$, we can decompose $x = x^+ - x^-$, where $x^+ = \max\{0, x\}$ and $x^- = \max\{0, -x\}$. It follows that $|x| = x^+ + x^-$. From this, we can write $b_j = b_j^+ - b_j^-$ and $r_j = r_j^+ - r_j^-$. Additionally, we introduce the following variables,

$$\begin{aligned} u_i &= \max\{0, Y_i - X_i'b - Z_i'r\} \\ v_i &= \max\{0, -Y_i + X_i'b + Z_i'r\} \\ \sigma_i &= \max\{0, |Y_i - X_i'b - Z_i'r| - \epsilon\} \\ s_i &= |Y_i - X_i'b - Z_i'r| - \epsilon - \sigma. \end{aligned}$$

The $n \times 1$ vectors u, v, σ, and s are obtained by stacking u_i, v_i, σ_i, and s_i, respectively, across all n observations. The first two constraints in (3) ensure that the decomposition of each term in (2) into the difference of its positive and negative components is consistent with the data. ■

B.2 Proof of Theorem II.1

Follows immediately from Lemma B.5, Lemma B.6, and Lemma B.7. ■

B.3 Derivation of Remark A.3

As shown in Lemma B.4,

$$n^{1/2}(\hat{\beta}_n - \beta) = \frac{1}{2}E[X_iX_i'f_Y(X_i'\beta - \epsilon|X_i, Z_i)]^{-1}n^{-1/2}\sum_{1\leq i\leq n} X_i(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\}) + o_P(1)$$

under Assumption I.1(a)–(e). Symmetry of F under Assumption I.2 implies that

$$n^{-1/2}\sum_{1\leq i\leq n} X_i(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\}) \xrightarrow{d} N(0, E[X_iX_i'(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\})^2]).$$

By the law of iterated expectations, the variance term may be expressed as

$$E[X_iX_i'(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\})^2] = E[X_iX_i'\ E[(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\})^2 \mid X_i, Z_i]].$$

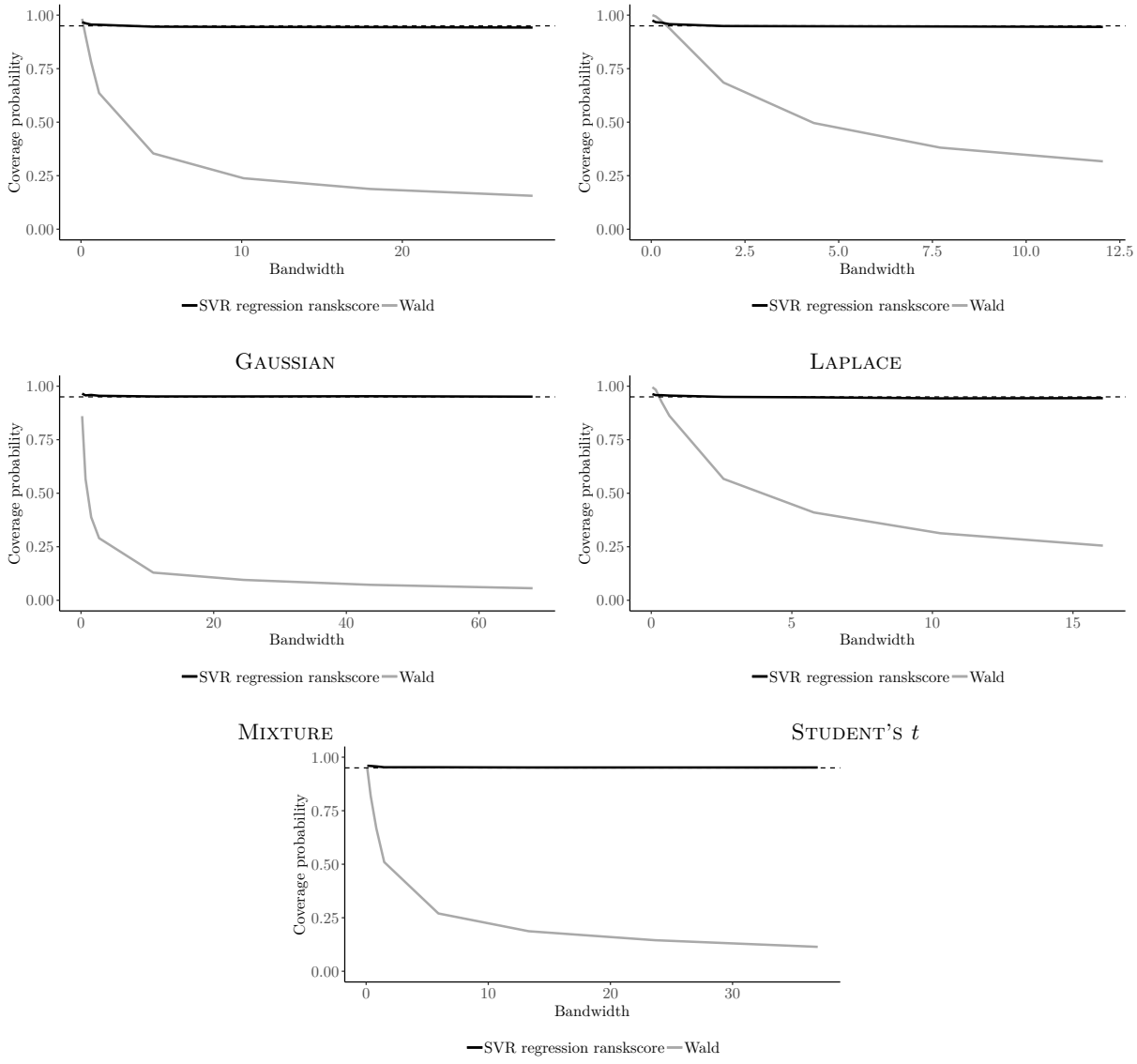

Figure 6: ℓ_1-SVR Regression Rankscore 95% Confidence Intervals versus 95% Wald Confidence Intervals

Note that

$$\begin{aligned}
&E[(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\})^2 | X_i, Z_i] \\
&= E[I\{Y_i \geq X_i'\beta + \epsilon\} + I\{Y_i \leq X_i'\beta - \epsilon\} | X_i, Z_i] \\
&= 2E[F_Y(X_i'\beta - \epsilon | X_i, Z_i)] \ ,
\end{aligned}$$

where the last equality follows from Assumption I.1(c). Substituting this into the expression for the asymptotic variance completes the derivation. ■

B.4 Proof of Corollary II.2

Follows immediately from Theorem II.1 and the duality between hypotheses tests and confidence regions. ■

B.5 Proof of Lemma A.1

By complementary slackness,

$$\begin{aligned}
|Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n| < \epsilon \Leftrightarrow s_i > 0 \Rightarrow \hat{a}_i^- = 0 \\
|Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n| > \epsilon \Leftrightarrow \sigma_i > 0 \Rightarrow \hat{a}_i^- = -1 \\
Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n > 0 \Leftrightarrow u_i > 0 \Rightarrow \hat{a}_i^+ + \hat{a}_i^- = 0 \\
Y_i - Z_i'\gamma_0 - X_i'\hat{\beta}_n < 0 \Leftrightarrow v_i > 0 \Rightarrow \hat{a}_i^- - \hat{a}_i^+ = 0 \ ,
\end{aligned}$$

and the result follows. ■

B.6 Proof of Theorem A.1

First note that the denominator of $T_n(\mathbf{W}_n, \gamma_0)$ is monotonically increasing in γ_0 because $\hat{p}_n$ in (9) is so. Next, we show the numerator is monotonically decreasing in γ_0. Denote by $\hat{a}^+(r)$ the solution to (6) with $\gamma_0 = r$. Given $r_1 > r_2$, by definition of the optimization problem,

$$(\mathbf{Y}_n - r_1\mathbf{Z}_n)'\hat{a}^+(r_1) - (\mathbf{Y}_n - r_1\mathbf{Z}_n)'\hat{a}^+(r_2) > 0 \ ,$$

while

$$(\mathbf{Y}_n - r_2\mathbf{Z}_n)'\hat{a}^+(r_1) - (\mathbf{Y}_n - r_2\mathbf{Z}_n)'\hat{a}^+(r_2) < 0 \ .$$

The two observations indicate that

$$r_1\mathbf{Z}_n'(\hat{a}^+(r_2) - \hat{a}^+(r_1)) > 0 > r_2\mathbf{Z}_n'(\hat{a}^+(r_2) - \hat{a}^+(r_1)) \ ,$$

so that

$$(r_1 - r_2)\mathbf{Z}_n'(\hat{a}^+(r_2) - \hat{a}^+(r_1)) > 0 \ ,$$

and the result follows since $r_1 > r_2$.

B.7 Auxiliary Lemmas

Lemma B.1. *Suppose U_i, $1 \le i \le n$ are i.i.d. random variables where $E|U_i|^r < \infty$. Then*

$$n^{-1/r} \max_{1 \le i \le n} |U_i| \xrightarrow{P} 0 \ .$$

Proof of Lemma B.1. Note that for all $\eta > 0$,

$$\begin{aligned}
P\left\{n^{-1/r} \max_{1 \le i \le n} |U_i| > \eta\right\} &\le P\left\{\max_{1 \le i \le n} |U_i|^r > n\eta^r\right\} \\
&\le nP\{|U_i|^r > n\eta^r\} \le \frac{n}{n\eta^r} E[|U_i|^r I\{|U_i|^r > n\eta^r\}] = \frac{1}{\eta^r} E[|U_i|^r I\{|U_i|^r > n\eta^r\}] \to 0 \ ,
\end{aligned}$$

where the convergence follows from the dominated convergence theorem and $E|U_i|^r \to 0$. ■

Lemma B.2. *Suppose P satisfies Assumption I.1(b)–(d). Then*

$$S(b) = E[\max\{|Y - X'b - \epsilon|, 0\}]$$

is uniquely minimized at $b = \beta$.

Proof of Lemma B.2. Follows immediately upon noting that Theorem 9.8 of Steinwart and Christmann (2008) holds under Assumption I.1(b)–(d). ■

Lemma B.3. *Suppose P satisfies Assumption I.1(b)–(d) and λ_n satisfies Assumption I.3. Then,*

$$\hat{\beta}_n \xrightarrow{P} \beta \ .$$

Proof of Lemma B.3. Define

$$S_n(b) = n^{-1} \sum_{1 \le i \le n} \max\{0, |Y_i - X_i'b| - \epsilon\} + \lambda_n \|b\|_1 \ .$$

To begin with, note that $S_n(b)$ is convex in b. Without any loss of generality suppose $\beta = 0$. For any $\delta > 0$, let B_δ denote the closed δ-ball around 0. By definition,

$$S_n(\hat{\beta}_n) \le S_n(0) \ . \tag{13}$$

For all $b \in \mathbb{R}^{d_x} \backslash B_\delta$, we have by convexity that

$$S_n(b_\delta) \le \frac{\delta}{|b|} S_n(b) + \frac{|b| - \delta}{|b|} S_n(0) \ ,$$

where

$$b_\delta = \frac{\delta}{|b|} b \ .$$

Therefore

$$S_n(b) \ge \frac{|b|}{\delta} S_n(b_\delta) - \frac{|b| - \delta}{\delta} S_n(0) \ . \tag{14}$$

Since $\{b : |b| = 1\}$ is compact and $S(b)$ is continuous in b, we have by Lemma B.2 that there exists $\eta > 0$ such that

$$\min_{b:|b|=1} S(b) \ge S(0) + \eta \ . \tag{15}$$

By Lemma 2.6.18 of Van Der Vaart and Wellner (1996), $\{b \to |y - x'b| - \epsilon + \lambda_n \|b\|_1 : |b| = 1\}$ is a VC class, thus Donsker, and thus Glivenko-Cantelli, i.e.,

$$\sup_{b:|b|=1} |S_n(b) - S(b)| = o_P(1) \ . \tag{16}$$

Combining (14), (15), and (16), and that $S_n(0) \xrightarrow{P} S(0)$, we have that

$$\{|\hat{\beta}_n - \beta| > \delta\} \Rightarrow \{S_n(\hat{\beta}_n) \ge S_n(0) + \eta + o_P(1)\} \ ,$$

which has probability approaching zero because of (13). ■

Lemma B.4. *Suppose P satisfies Assumption I.1(a)–(e). Then,*

$$n^{1/2}(\hat{\beta}_n - \beta) = \frac{1}{2} E[X_i X_i' f_Y(X_i'\beta - \epsilon | X_i, Z_i)]^{-1} n^{-1/2} \sum_{1 \le i \le n} X_i (I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\}) + o_P(1) \ .$$

Proof of Lemma B.4. Define

$$\hat{L}_n = n^{-1/2} \sum_{1 \le i \le n} X_i (I\{Y_i - X_i'\hat{\beta}_n > \epsilon\} - I\{Y_i - X_i'\hat{\beta}_n < -\epsilon\}) \tag{17}$$

$$L_n = n^{-1/2} \sum_{1 \le i \le n} X_i (I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\}) \ . \tag{18}$$

To begin with, note that

$$\left| n^{-1/2} \sum_{1 \le i \le n} X_i \hat{a}_i^+ - \hat{L}_n \right| \lesssim n^{-1/2} \max_{1 \le i \le n} |X_i| \sum_{1 \le i \le n} (I\{Y_i = X_i'\hat{\beta}_n\} + I\{Y_i - X_i'\hat{\beta}_n = \epsilon\}) = o_P(1) \ , \tag{19}$$

because of by Lemma A.1, Lemma B.1, Assumption I.1(a), and that the number of support vectors are bounded. By similar arguments,

$$\left| n^{-1/2} \sum_{1 \le i \le n} X_i \hat{a}_i^+ \right| \le \left| n^{-1/2} \lambda_n \mathbf{1}_{d_x} \right| + o_P(1) = o_P(1) \ , \tag{20}$$

where the second equality follows from (6) and the last follows from Lemma B.1. Next, we write

$$\begin{aligned}\hat{L}_n - L_n &= n^{-1/2} \sum_{1\le i\le n} X_i(I\{Y_i \le X_i'\beta + \epsilon\} - I\{Y_i \le X_i'\hat{\beta}_n + \epsilon\}) \\ &\quad + n^{-1/2} \sum_{1\le i\le n} X_i(I\{Y_i \le X_i'\beta - \epsilon\} - I\{Y_i \le X_i'\hat{\beta}_n - \epsilon\}) \qquad (21)\\ &= R_{1,n} + R_{2,n} + R_{1,n}^- + R_{2,n}^- \ ,\end{aligned}$$

where

$$\begin{aligned}R_{1,n} &= n^{-1/2} \sum_{1\le i\le n} X_i I\{Y_i \le X_i'\beta + \epsilon\} - E[X_i I\{Y_i \le X_i'\beta + \epsilon\}] \\ &\quad - (X_i I\{Y_i \le X_i'\hat{\beta}_n + X_i't + \epsilon\} - E[X_i I\{Y_i \le X_i'\beta + X_i't + \epsilon\}])|_{t=\hat{\beta}_n-\beta} \\ R_{2,n} &= n^{1/2}E[X_i(I\{Y_i \le X_i'\beta + \epsilon\} - I\{Y_i \le X_i'\beta + n^{-1/2}X_i't + \epsilon\})]|_{t=n^{1/2}(\hat{\beta}_n-\beta)} \ ,\end{aligned}$$

and similarly for $R_{1,n}^-$ and $R_{2,n}^-$.

Since $\hat{\beta}_n - \beta = o_P(1)$ by Lemma B.3, $R_{1,n} \xrightarrow{P} 0$ by the similar arguments as those used in the last part of the proof of Lemma A.1 of Bai, Pouliot and Shaikh (2019). For $R_{2,n}$, note that

$$\begin{aligned}R_{2,n} &= n^{1/2}E[X_i(F_Y(X_i'\beta + \epsilon|X_i, Z_i) - F_Y(X_i'\beta + n^{-1/2}X_i't + \epsilon|X_i, Z_i))]|_{t=n^{1/2}(\hat{\beta}_n-\beta)} \\ &= -n^{1/2}E[X_i n^{-1/2}X_i' t f_Y(X_i'\beta + \epsilon + s_i n^{-1/2}X_i't|X_i, Z_i)]|_{t=n^{1/2}(\hat{\beta}_n-\beta)} \\ &= -n^{1/2}(\hat{\beta}_n - \beta)E[X_i X_i' f_Y(X_i'\beta + \epsilon + s_i n^{-1/2}X_i't|X_i, Z_i)]|_{t=n^{1/2}(\hat{\beta}_n-\beta)} \qquad (22)\end{aligned}$$

where $s_i \in [0,1]$ is a random variable. The first equality above holds by the law of iterated expectation and the second holds by the mean-value theorem. A similar decomposition holds for $R_{1,n}^-$ and $R_{2,n}^-$.

We then argue that $L_n = O_P(1)$. Indeed, by Assumption I.1(c), the conditional distributions are symmetric so that the individual terms of L_n are i.i.d. mean zero and therefore $L_n = O_P(1)$ by the central limit theorem.

By applying similar arguments as those used to establish Lemma A.2 of Bai, Pouliot and Shaikh (2019), where assumptions are satisfied under Assumption I.1(a)–(e), and noting that $L_n = O_p(1)$, it follows from (17), (18), (19), (20), and (22) that

$$n^{1/2}(\hat{\beta}_n - \beta) = O_p(1) \ . \qquad (23)$$

and

$$n^{1/2}(\hat{\beta}_n - \beta)(E[X_i X_i'\big(f_Y(X_i'\beta + \epsilon|X_i, Z_i) + f_Y(X_i'\beta - \epsilon|X_i, Z_i)\big)] + o_P(1)) = -L_n + o_P(1) \ . \qquad (24)$$

The proof is finished by plugging (23) in (24), and noting that

$$f_Y(X_i'\beta + \epsilon|X_i, Z_i) = f_Y(X_i'\beta - \epsilon|X_i, Z_i)$$

by Assumption I.1(c). ■

Lemma B.5. *Suppose P satisfies Assumption I.1(a)–(d) and ϵ and λ_n satisfies Assumption I.3. Then,*

$$n^{-1/2}\mathbf{Z}_n'\hat{a}^+ \xrightarrow{d} N(0, 2E[\tilde{Z}_i\tilde{Z}_i' F_Y(X_i'\beta - \epsilon|X_i, Z_i)]) \ ,$$

where

$$\tilde{Z}_i = Z_i - E[Z_i X_i' f_Y(X_i'\beta - \epsilon|X_i, Z_i)]E[X_i X_i' f_Y(X_i'\beta - \epsilon|X_i, Z_i)]^{-1}X_i$$

Proof of Lemma B.5. It follows from Lemma B.3, Lemma B.4, and similar arguments used to establish Lemma A.1 of Bai, Pouliot and Shaikh (2019) that

$$n^{-1/2}\mathbf{Z}_n'\hat{a}^+ = n^{-1/2} \sum_{1\le i\le n} \tilde{Z}_i(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\}) + o_P(1) \ .$$

Note that

$$\begin{aligned}
&E[(I\{Y_i - X_i'\beta > \epsilon\} - I\{Y_i - X_i'\beta < -\epsilon\})^2 | X_i, Z_i] \\
&= E[I\{Y_i \geq X_i'\beta + \epsilon\} + I\{Y_i \leq X_i'\beta - \epsilon\} | X_i, Z_i] \\
&= 2E[F_Y(X_i'\beta - \epsilon | X_i, Z_i)] \ ,
\end{aligned}$$

where the last equality follows from Assumption I.1(c). The lemma now follows from the Central Limit Theorem and Assumption I.1(a). ■

Lemma B.6. *Suppose P satisfies Assumption I.1(a) and Assumption I.2(a). Then,*

$$n^{-1}\mathbf{Z}_n'\mathbf{M}_n\mathbf{Z}_n \xrightarrow{P} E[Z_iZ_i'] - E[Z_iX_i']E[X_iX_i']^{-1}E[X_iZ_i'] \ .$$

Proof of Lemma B.6. Follows from Assumption I.1(a), Assumption I.2(a), and an application of the weak law of large numbers. ■

Lemma B.7. *Suppose P satisfies Assumptions I.1(a)–(d) and I.2, and λ_n satisfies Assumption I.3. Then,*

$$\hat{p}_n \xrightarrow{P} 2p_\epsilon \ .$$

Proof of Lemma B.7. We consider

$$\frac{1}{n}\sum_{1\leq i\leq n} I\{Y_i \leq X_i'\beta + Z_i'\gamma_0 - \epsilon + X_i'(\hat{\beta}_n - \beta)\} \ ,$$

and the other half follows similarly. By Lemma B.3, since Assumptions I.1(a)–(d) and I.3 hold, $\hat{\beta}_n \xrightarrow{P} \beta$. Fix $\eta > 0$. For any $\delta > 0$, consider the empirical process indexed by the class of functions

$$\{t \to I\{y \leq x'\beta + z'\gamma_0 - \epsilon + x't\} : \|t\| \in [0, \delta]\} \ .$$

It is easy to see the class of functions is VC by Lemma 9.12 of Kosorok (2008), so that is Donsker hence Glivenko-Cantelli by Theorem 2.6.7 of Van Der Vaart and Wellner (1996), i.e.,

$$\sup_{t\in[0,\delta]} \left| \frac{1}{n}\sum_{1\leq i\leq n} I\{Y_i \leq X_i'\beta + Z_i'\gamma_0 - \epsilon + X_i't\} - P\{Y_i \leq X_i'\beta + Z_i'\gamma_0 - \epsilon + X_i't\} \right| \xrightarrow{P} 0 \ .$$

Next,

$$P\{Y_i \leq X_i'\beta + Z_i'\gamma_0 - \epsilon + X_i't\}$$

is continuous at $t = 0$. Since $\hat{\beta}_n - \beta = o_P(1)$, with probability approaching 1, $\|\hat{\beta}_n - \beta\| \leq \delta$, and therefore

$$\left| \frac{1}{n}\sum_{1\leq i\leq n} I\{Y_i \leq X_i'\beta + Z_i'\gamma_0 - \epsilon + X_i't\} - P\{Y_i \leq X_i'\beta + Z_i'\gamma_0 - \epsilon\} \right| \leq \eta + \eta_\delta \ ,$$

where $\eta_\delta \to 0$ as $\delta \to 0$. Let $\delta \to 0$ to finish the proof. ■

References

Bai, Yuehao, Guillaume A Pouliot, and Azeem M Shaikh. 2019. "On Regression Rankscore Inference." Working Paper.

Kosorok, Michael R. 2008. *Introduction to empirical processes and semiparametric inference.* Springer.

Pouliot, Guillaume A. 2020. "Instrumental Variables Quantile Regression with Multivariate Endogenous Variable." Working Paper.

Powell, James L. 1991. "Estimation of monotonic regression models under quantile restrictions." *Nonparametric and semiparametric methods in Econometrics*, 357–384.

Steinwart, Ingo, and Andreas Christmann. 2008. *Support vector machines.* Springer Science & Business Media.

Van Der Vaart, Aad W, and Jon A Wellner. 1996. "Weak convergence." In *Weak convergence and empirical processes.* 16–28. Springer.