AEA Papers and Proceedings 2021, 111: 611–615
https://doi.org/10.1257/pandp.20211035

OPTIMIZATION-CONSCIOUS ECONOMETRICS[‡]

Inference for Support Vector Regression under ℓ_1 Regularization[†]

By Yuehao Bai, Hung Ho, Guillaume A. Pouliot, and Joshua Shea*

This paper studies inference for support vector regression (SVR) with ℓ_1-norm regularization (ℓ_1-SVR). SVR is the extension of the support vector machine (SVM) classification method (Vapnik 1998) to the regression problem (Basak, Pal, and Patranabis 2007) and is designed to reproduce the good out-of-sample performance of SVM classification in the regression setting. It has been frequently used in regression analysis across fields such as geophysical sciences (Ghorbani, Zargar, and Jazayeri-Rad 2016), engineering (Li, West, and Platt 2012), and image compression (Jiao et al. 2005).

However, theory and closed-form expressions for the asymptotic variance of the regression coefficient estimates, or of tests that may be inverted for inference, are not available. While there is nonasymptotic methodology for inference, it is limited to small samples and relies on distributional assumptions (Gao et al. 2002; Law and Kwok 2001). Because such assumptions are typically not satisfied in practice, we find these methods impractical.

It has been shown that calculations akin to those used to derive asymptotic distributions of quantile regression coefficient estimates may be used to produce asymptotic approximations of conditional (on features) probabilities of classification for SVM (Pouliot 2018). These derivations may be extended to produce the asymptotic distribution of the regression coefficients in SVR, but they rely on a nonparametric estimate of the density of the regression errors. Density estimation itself requires an arbitrary choice of bandwidth parameter and may allow users to present deceptively narrow confidence intervals whose coverage properties fall well below the nominal level. See Figure 1.

This paper addresses these issues and delivers, to the best of our knowledge, the first derivation of error bars for SVR that does not require distributional assumptions and the first rigorous treatment of large-sample inference for SVR. We further improve on this by developing a bandwidth-free procedure based on the inversion of a novel regression rank score test statistic that displays competitive power properties.[1]

I. Setup and Notation

Let $W_i = (Y_i, X_i, Z_i) \in \mathbb{R} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}$, $1 \leq i \leq n$ be i.i.d. random vectors. We assume the first element of X_i is 1. Let P denote the distribution of W_i. For a random variable (vector) A, define the vector (matrix) $\mathbf{A}_n = (A_1, \ldots, A_n)'$. Let $Q_Y(x,z)$ denote the conditional median of Y given $X = x, Z = z$. We assume that this regression function is linear, that is,

$$Q_Y(x,z) = x'\beta(P) + z'\gamma(P), \tag{1}$$

where $\beta(P) \in \mathbb{R}^{d_x}$ and $\gamma(P) \in \mathbb{R}^{d_z}$ are unknown parameters. We omit the dependence of β and γ on P whenever it is clear from the context.

[‡] *Discussants:* Elie Tamer, Harvard University; Francesca Molinari, Cornell University; Alexander Torgovitsky, University of Chicago; Xiaohong Chen, Yale University.

* Bai: Department of Economics, University of Michigan (email: yuehaob@umich.edu); Ho: The Wharton School, University of Pennsylvania (email: hqdh@wharton.upenn.edu); Pouliot: Harris School of Public Policy, University of Chicago (email: guillaumepouliot@uchicago.edu); Shea: Department of Economics, University of Chicago (email: jkcshea@uchicago.edu).

[†] Go to https://doi.org/10.1257/pandp.20211035 to visit the article page for additional materials and author disclosure statement(s).

[1] An R package for estimation and inference using ℓ_1-SVR is available at https://github.com/jkcshea/l1svr.

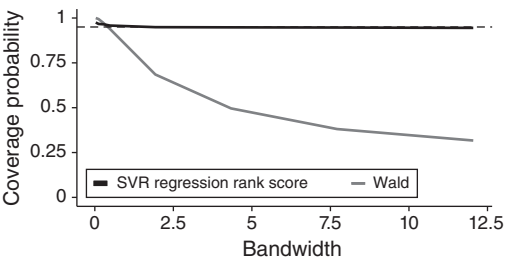

FIGURE 1. ℓ_1-SVR REGRESSION RANK SCORE CONFIDENCE INTERVAL VERSUS WALD CONFIDENCE INTERVAL

Notes: Simulated coverage probabilities of the 95 percent confidence interval when errors are heteroskedastic and Laplacian. These results extend to all other heteroskedastic error distributions considered in Table 1.

The covariates are (X_i, Z_i). We distinguish X_i and Z_i to make transparent that the covariate Z_i is the one for which we conduct inference.

Consider the following ℓ_1-SVR:

$$(2)\quad \min_{(b,r)\in \mathbb{R}^{d_x}\times\mathbb{R}^{d_z}} n^{-1}\sum_{1\le i\le n}\max\{0,|Y_i - X_i'b - Z_i'r| - \epsilon\} + \lambda_n(\|b\|_1 + \|r\|_1),$$

where

$$\|b\|_1 = \sum_{1\le j\le d_x}|b_j|$$

and similarly for $\|r\|_1$.

Define $F_Y(y|x,z)$ as the conditional distribution at $Y = y$ given $X = x$ and $Z = z$ and $f_Y(y|x,z)$ as the corresponding conditional density. We impose the following conditions on the distribution P.

ASSUMPTION 1: *The distribution P is such that*

(*i*) $E\left[\begin{pmatrix} X_iX_i' & X_iZ_i' \\ Z_iX_i' & Z_iZ_i'\end{pmatrix} f_Y(X_i'\beta + Z_i'\gamma - \epsilon | X_i, Z_i)\right]$ *is strictly positive and finite.*

(*ii*) $f_Y(y|x,z)$ *exists for all* $(y,x,z) \in \mathbb{R}\times\mathbb{R}^{d_x}\times\mathbb{R}^{d_z}$.

(*iii*) $f_Y(\cdot\,|x,z)$ *is symmetric around* $x'\beta + z'\gamma$ *for all* $(x,z) \in \mathbb{R}^{d_x}\times\mathbb{R}^{d_z}$.

(*iv*) $f_Y(x'\beta + z'\gamma - \epsilon\,|x,z) > 0$ *for all* $(x,z) \in \mathbb{R}^{d_x}\times\mathbb{R}^{d_z}$.

(*v*) *Define*

$$\Gamma = \{(x,z) \in \mathbb{R}^{d_x}\times\mathbb{R}^{d_z}: y \in [x'\beta + z'\gamma - c, x'\beta + z'\gamma + c]\}.$$

There exists $c > 0$ *such that*

$$\sup_{(x,z)\in\Gamma}\frac{|f_Y(y\,|x,z) - f_Y(x'\beta + z'\gamma\,|x,z)|}{|y - x'\beta - z'\gamma|} < \infty.$$

Assumption 1 (i), (ii), and (v) are commonly imposed in the quantile regression literature in order to establish the asymptotic distributions of estimators (Koenker 2005; Bai, Pouliot, and Shaikh 2019). Assumption 1 (iii)–(iv) are imposed so that the coefficient estimates from the ℓ_1-SVR model are consistent for the coefficients of the linear conditional median regression function.

For pivotal inference, we will require the following strong but powerful homoskedasticity assumption on P.

ASSUMPTION 2: *The distribution P is such that*

(*i*) $f_Y(x'\beta + z'\gamma - \epsilon\,|x,z) = g(\epsilon)$ *for all* $(x,z) \in \mathbb{R}^{d_x}\times\mathbb{R}^{d_z}$, *for some function* g.

(*ii*) $F_Y(x'\beta + z'\gamma - \epsilon\,|x,z) \equiv p_\epsilon$ *across all* $(x,z) \in \mathbb{R}^{d_x}\times\mathbb{R}^{d_z}$, *where the constant* p_ϵ *may depend on* ϵ.

Assumption 2 (i) is imposed so that the density terms cancel in the expression of the limiting variances of the test statistic, thus delivering pivotal inference. Assumption 2 (ii) is imposed so that the test statistic is simpler but is not required in general.

As in Ghorbani, Zargar, and Jazayeri-Rad (2016), we impose the following condition on the tuning parameter λ_n. It is satisfied when $\lambda_n = \lambda/n$, where λ is a constant.

ASSUMPTION 3: $\lambda_n \to 0$ *as* $n \to \infty$.

Let $\mathbf{1}_d$ denote the $d \times 1$ vector of ones. The ℓ_1-SVR problem (2) has the following primal linear programming formulation:

$$\max \mathbf{1}_n' \sigma + \lambda_n\big(\mathbf{1}_{d_x}' b^+ + \mathbf{1}_{d_x}' b^- + \mathbf{1}_{d_z}' r^+ + \mathbf{1}_{d_z}' r^-\big), \tag{3}$$

subject to

$$\begin{aligned} u - v &= \mathbf{Y}_n - \mathbf{Z}_n r - \mathbf{X}_n b, \\ \sigma - s &= u + v - \epsilon \mathbf{1}_n, \\ b^+, b^-, r^+, r^-, u, v, \sigma, s &\geq 0, \end{aligned}$$

where the optimization is over b^+, b^-, r^+, r^-, u, v, σ, s. See the online Appendix for the derivation of (3).

II. Inference

For a prespecified $\gamma_0 \in \mathbb{R}^{d_z}$, we are interested in inverting tests of

$$H_0: \gamma(P) = \gamma_0 \quad \text{versus} \quad H_1: \gamma(P) \neq \gamma_0 \tag{4}$$

at level $\alpha \in (0,1)$.

For that purpose, consider the following "short" ℓ_1-SVR problem constructed by replacing $\mathbf{Y}_n$ with $\mathbf{Y}_n - \mathbf{Z}_n\gamma_0$ and omitting $\mathbf{Z}_n$ from the regressors in (3):

$$\max_{b^+, b^-, u, v, \sigma, s} \mathbf{1}_n' \sigma + \lambda_n\big(\mathbf{1}_{d_x}' b^+ + \mathbf{1}_{d_x}' b^-\big) \tag{5}$$

subject to

$$\begin{aligned} u - v &= \mathbf{Y}_n - \mathbf{Z}_n\gamma_0 - \mathbf{X}_n\big(b^+ - b^-\big), \\ \sigma - s &= u + v - \epsilon \mathbf{1}_n, \\ b^+, b^-, u, v, \sigma, s &\geq 0. \end{aligned}$$

Define $\hat{\beta}_n$ as $b^+ - b^-$, where b^+ and b^- are part of the solution to (5). The dual of (5) is

$$\max_{a^+, a^-} \big(\mathbf{Y}_n - \mathbf{Z}_n\gamma_0\big)' a^+ + \epsilon \mathbf{1}_n' a^- \tag{6}$$

subject to

$$\begin{aligned} -\lambda_n \mathbf{1}_{d_x} &\leq \mathbf{X}_n' a^+ \leq \lambda_n \mathbf{1}_{d_x}, \\ a^- &\leq a^+ \leq -a^-, \\ a^- &\in [-1, 0]^n. \end{aligned}$$

Denote the solution to (6) by $\hat{a}^+$ and $\hat{a}^-$.

We construct the *ℓ_1-SVR regression rank score test statistic* as

$$T_n\big(\mathbf{W}_n, \gamma_0\big) = \frac{n^{-1/2}\mathbf{Z}_n'\hat{a}^+}{\sqrt{n^{-1}\mathbf{Z}_n'\mathbf{M}_n\mathbf{Z}_n\hat{p}_n}}, \tag{7}$$

$$\mathbf{M}_n = I - \mathbf{X}_n\big(\mathbf{X}_n'\mathbf{X}_n\big)^{-1}\mathbf{X}_n', \tag{8}$$

and

$$\hat{p}_n = \frac{1}{n}\sum_{1 \leq i \leq n} I\Big\{\big|Y_i - X_i'\hat{\beta}_n - Z_i'\gamma_0\big| \geq \epsilon\Big\}. \tag{9}$$

We define the *ℓ_1-SVR regression rank score test* as

$$\phi_n\big(\mathbf{W}_n, \gamma_0\big) = I\Big\{\big|T_n\big(\mathbf{W}_n, \gamma_0\big)\big| > z_{1-\frac{\alpha}{2}}\Big\}, \tag{10}$$

where $z_{1-\frac{\alpha}{2}}$ is the $\big(1 - \alpha/2\big)$th quantile of the standard normal distribution.

To see the intuition behind the construction of $T_n\big(\mathbf{W}_n, \gamma_0\big)$, consider running the SVR of $Y - Z'\gamma_0$ on X, with $\hat{a}^+$ as the solution to the dual problem. If H_0 holds, that is, $\gamma(P) = \gamma_0$, then regressing $Y - Z'\gamma_0$ on X and Z should result in an estimated coefficient "close" to 0 on Z. Hence, whether or not Z is included in the regression should have "close" to zero effect on the primal or dual results. Equivalently, adding the constraint $\mathbf{Z}_n'\hat{a}^+ = 0$ to (6) should not change the solution very much, so that $\mathbf{Z}_n'\hat{a}^+ = 0$ holds approximately when the null hypothesis holds but may be large otherwise.

The following theorem is our main result. It establishes the asymptotic distribution of the test statistic defined in (7) under the null and guarantees asymptotic exactness of the test defined in (10).

THEOREM 1: *Suppose P satisfies Assumption 1, λ_n satisfies Assumption 3, and P additionally satisfies the null hypothesis, that is, $\gamma(P) = \gamma_0$. Then,*

$$n^{-1/2}\mathbf{Z}_n'\hat{a}^+ \xrightarrow{d} N\Big(0, 2E\Big[\tilde{Z}_i\tilde{Z}_i' F_Y\big(X_i'\beta + Z_i'\gamma_0 - \epsilon | X_i, Z_i\big)\Big]\Big), \tag{11}$$

where

$$\tilde{Z}_i = Z_i - E\Big[Z_i X_i' f_Y\big(X_i'\beta + Z_i'\gamma_0 - \epsilon | X_i, Z_i\big)\Big] \times E\Big[X_i X_i' f_Y\big(X_i'\beta + Z_i'\gamma_0 - \epsilon | X_i, Z_i\big)\Big]^{-1} X_i.$$

If P additionally satisfies Assumption 2, then

$$T_n(\mathbf{W}_n, \gamma_0) \xrightarrow{d} N(0,1),$$

and therefore, for the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n(\mathbf{W}_n)$ *defined in (10) satisfies*

$$\lim_{n\to\infty} E\big[\phi_n(\mathbf{W}_n, \gamma_0)\big] = \alpha.$$

Moreover, the following corollary delivers pivotal inference and allows the test to be performed easily.

COROLLARY 1: *Suppose P satisfies Assumptions 1–2 and the null hypothesis and* λ_n *satisfies Assumption 3. Then the asymptotic variance in (11) can be consistently estimated without density estimation by*

$$\frac{1}{n}\mathbf{Z}_n'\mathbf{M}_n\mathbf{Z}_n\hat{p}_n,$$

where $\mathbf{M}_n$ *and* $\hat{p}_n$ *are defined in (8) and (9), respectively.*

According to Theorem 1, we can construct confidence regions by inverting the test $\phi_n(\mathbf{W}_n, \gamma_0)$ in (10). The following corollary shows that the limiting coverage probability of the confidence region is indeed correct.

COROLLARY 2: *Let* $\phi_n(\mathbf{W}_n, \gamma_0)$ *denote the test in (10) with level* α. *Define*

$$(12)\quad C_n = \big\{\gamma_0 \in \mathbb{R} : \phi_n(\mathbf{W}_n, \gamma_0) = 0\big\}.$$

Suppose P satisfies Assumptions 1 and 2 and λ_n *satisfies Assumption 3. Then*,

$$\lim_{n\to\infty} P\{\gamma \in C_n\} = 1 - \alpha.$$

We show in the online Appendix that the test statistic is monotonic, which guarantees the confidence region is an interval.

III. Simulation

This section presents a simulation study on the size, power, and width of the error bars for the ℓ_1-SVR regression rank score test in finite samples. We compare its performance against the median regression rank score test, a natural benchmark (Koenker 2005; Bai, Pouliot, and Shaikh 2019).

The data-generating process is

$$Y = -0.8 + 2X + \gamma Z + G^{-1}(\tau),$$

$$(X, Z) \sim N\left(\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}10 & 4\\4 & 8\end{pmatrix}\right),$$

where G^{-1} is the inverse CDF for the error, τ is uniformly distributed over the $[0,1]$ interval, and $(X, Z) \perp\!\!\!\perp \tau$. In all simulations, the sample size is 500 and we set $\lambda_n = 0$. The parameter ϵ is adjusted according to the distribution of the error so that $G(\epsilon) - G(-\epsilon) = 0.2$. The results of the simulation carry over to cases of higher dimensional vectors of covariates so long as the sample size is sufficiently large.

Table 1 presents the simulation results for four distributions of error terms: Gaussian, a symmetric mixture of Gaussian distributions, Student's t, and χ^2. The latter three distributions allow us to measure the performance of the test when the error distribution exhibits either multiple modes, fat tails, or asymmetry. We set the true parameter $\gamma = 0$ under the null to study size properties and $\gamma = 0.5$ under the alternative to study power properties at critical level $\alpha = 0.05$. Additional simulations and their details may be found in the online Appendix.

Rows 1–2 indicate that the size properties of the SVR and median regression rank score tests are about equal under homoskedasticity, but the SVR regression rank score test often has better power properties. Rows 3–4 reiterate these findings in the case where the errors are heteroskedastic. Rows 5–6 likewise show that the confidence intervals under homoskedasticity obtained through the SVR regression rank score test are often narrower than those of the median regression rank score test. Additional simulations in the online Appendix yield similar comparisons between the two inference methods.

IV. Conclusion

In this article, we developed classical large-sample inference and furthermore delivered methodology producing asymptotically valid error bars while circumventing the need to select a bandwidth parameter. The asymptotic theory developed to establish the validity

TABLE 1—REJECTION PROBABILITIES AND CONFIDENCE INTERVALS FOR DIFFERENT DISTRIBUTIONS OF THE ERRORS

	Gaussian		Mixture		Student's t		χ^2	
	SVR	QR	SVR	QR	SVR	QR	SVR	QR
Homoskedasticity size, $\gamma = 0.0$	5.7	5.4	4.6	4.3	5.7	5.4	5.1	5.4
Homoskedasticity power, $\gamma = 0.5$	34.6	31.3	37.9	30.6	40.1	35.9	39.3	39.6
Heteroskedasticity size, $\gamma = 0.0$	4.5	4.7	4.5	4.4	4.4	4.6	4.7	3.7
Heteroskedasticity power, $\gamma = 0.5$	34.9	30.3	37.7	32.1	37.0	31.7	16.4	15.4
Homoskedasticity 95 percent CI, lower	0.03	−0.02	0.01	−0.03	0.07	0.03	0.07	0.06
Homoskedasticity 95 percent CI, upper	1.02	1.05	1.00	1.05	0.97	1.00	0.96	0.97

of the error bars is novel for SVR and may be of independent interest. Remarkably, simulation evidence suggests that the regression rank score test with our proposed regression rank score test statistic may outperform the standard median regression rank score test in inference for the regression parameters of the linear median regression function.

REFERENCES

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

Basak, Debasish, Srimanta Pal, and Dipak Chandra Patranabis. 2007. "Support Vector Regression." *Neural Information Processing - Letters and Reviews* 11 (10): 203–24.

Gao, Junbin B., Steve R. Gunn, Chris J. Harris, and Martin Brown. 2002. "A Probabilistic Framework for SVM Regression and Error Bar Estimation." *Machine Learning* 46 (1–3): 71–89.

Ghorbani, Mohammad, Ghasem Zargar, and Hooshang Jazayeri-Rad. 2016. "Prediction of Asphaltene Precipitation Using Support Vector Regression Tuned with Genetic Algorithms." *Petroleum* 2 (3): 301–06.

Jiao, Runhai, Yuancheng Li, Qingyuan Wang, and Bo Li. 2005. "SVM Regression and Its Application to Image Compression." In *Advances in Intelligent Computing: International Conference on Intelligence Computing, ICI 2005, Hefei, China, August 23–26, 2005, Proceedings, Part I. Vol. 3644 of the Lecture Notes in Computer Science*, edited by De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, 747–56. Cham, Switzerland: Springer.

Koenker, Roger. 2005. *Quantile Regression*. Cambridge, UK: Cambridge University Press.

Law, Martin H.C., and James Tin-Yau Kwok. 2001. "Bayesian Support Vector Regression." *AISTATS Proceedings*. http://www.gatsby.ucl.ac.uk/aistats/aistats2001/files/law119.ps.

Li, Jiaming, Sam West, and Glenn Platt. 2012. "Power Decomposition Based on SVM Regression." In *2012 Proceedings of International Conference on Modelling, Identification and Control*, 1195–99. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Pouliot, Guillaume. 2018. "Equivalence of Multicategory SVM and Simplex Cone SVM: Fast Computations and Statistical Theory." In *Proceedings of the 35th International Conference on Machine Learning, PMLR*, Vol. 80, 133–40.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory.* New York: John Wiley & Sons Inc.