Efficient Prediction Intervals via Debiased Conformal Threshold Ridge Regression

Jiamei Wu[1], Pan Shang[1], Yanlin Tang[2], Linglong Kong[3], Bei Jiang[3] and Lingchen Kong[1]

[1]*School of Mathematical and Statistics, Beijing Jiaotong University*

[2] *KLATASDS-MOE, School of Statistics, East China Normal University*

[3] *Department of Mathematical and Statistical Sciences, University of Alberta*

Supplementary Material

S1 Additional Assumptions for Theorems

Besides the assumptions in Section 3.2, we need some additional assumptions for Theorem 2–4. For deterministic sequences, $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ denote the usual asymptotic upper and lower bounds. For random sequences, $a_n = O_p(b_n)$ means $|a_n|/b_n$ is bounded in probability, and $a_n = o_p(b_n)$ means $|a_n|/b_n \to 0$ in probability.

B1 The smallest positive singular value λ_r of $\mathbf{X}_{\text{train}}$ satisfies

$$\lambda_r = \Omega\left(n^{\eta}\right), \quad 0 < \eta \leq 1.$$

B2 $\|\theta\| = O\left(n^{\alpha_\theta}\right)$ with $0 < \alpha_\theta < 4\eta$ and $p = O\left(n^{\alpha_p}\right)$ with $0 < \alpha_p < \infty$.

B3 $\mathbb{E}|\epsilon_i|^m < \infty$ for an even integer m satisfying $\alpha_p < m\eta$.

B4 $h_n = O(n^{\delta})$ with $\delta < \frac{3\eta - \alpha_\theta}{2}$.

B5 $a_n = O(n^{-\alpha_a})$ with $0 < \alpha_a < \eta - \frac{\alpha_p}{m}$. Furthermore, we assume that there exists a constant $0 < c_a < 1$ such that $\max_{i \notin \mathcal{M}_{a_n}} |\theta_i| \leq c_a \times a_n$, and $\min_{i \in \mathcal{M}_{a_n}} |\theta_i| \geq \frac{a_n}{c_a}$, where $\mathcal{M}_{a_n} = \left\{ j \mid |\widetilde{\theta}_j| > a_n \right\}$.

B6 There exists a constant $C_{\mathcal{N}} > 0$ such that

$$\max_{1 \leq i \leq n} \sum_{j \in \mathcal{M}_{a_n}} x_{ij}^2 \leq C_{\mathcal{N}}.$$

These assumptions are standard in the high-dimensional statistical literature. Assumption B1 requires nondegeneracy of the column space of $\mathbf{X}_{\text{train}}$, ensuring stability to perturbations while still allowing strong correlations or multicollinearity; see Bai and Yin (1993) for examples. In high-dimensional regression contexts, the sparsity of β implies the sparsity of θ in terms of the L_2-norm; see Shao and Deng (2012) for a detailed discussion. Thus, we impose a sparsity condition on θ and require that the dimension of the parameter vector p diverges at a polynomial rate, which can be much larger than n as outlined in Assumption B2. Assumption B3 imposes a mild finite-moment condition on the noise distribution to control the effect of extreme values. A stronger but more restrictive alternative is to impose a normality condition (i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$), which is a special case satisfying

Assumption B3. Assumptions B4–B6 are analogous to those in Zhang and Politis (2022), but are stated here in a more relaxed form. Assumption B4 restricts the growth (or shrinkage) rate of the ridge regularization parameter h_n. The rate condition $\delta < (3\eta - \alpha_\theta)/2$ balances the regularization bias and variance, ensuring that bias does not dominate the limiting behavior of the estimator. Intuitively, when the signal grows faster (larger α_θ), the admissible δ must be smaller. Conversely, greater stability of the design (larger η) permits more aggressive regularization. Importantly, B4 places no restriction on the limit of h_n: it may shrink, remain bounded, or diverge, as long as the specified rate condition is satisfied. Assumptions B5 and B6 are technical conditions ensuring separation of nonzero coefficients and boundedness of transformed design terms, respectively.

S2 Lemmas and Proofs

Lemma 1. *For every $\theta \in \mathbb{R}^d$ and $t > 0$, define $\mathrm{F}_\theta(t) = \mathbb{P}(|Y - X\theta| \leq t)$ as the cumulative distribution function and $\mathrm{F}_\theta^{-1}(p) = \inf\{t \in \mathbb{R} : \mathrm{F}_\theta(t) \geq p\}$ as the quantile function of the $|Y - X\theta|$ respectively. If the density of $|Y - X\theta|$ is bounded below by $\ell > 0$ in a neighborhood of its α upper quantile, then F_θ^{-1} is 1-Hölder continuous on this neigborhood with Hölder continuity constant*

$1/\ell$, *i.e.*,

$$|\mathrm{F}_\theta^{-1}(q_2) - \mathrm{F}_\theta^{-1}(q_1)| \;\leq\; \tfrac{1}{\ell}|q_2 - q_1|.$$

Proof. Without loss of generality, for any $q_2 \geq q_1$ which are in the neighborhood of the α upper quantile, denote $\mathbb{P}\left(|Y - X\theta| \leq t_1\right) = q_1, \mathbb{P}\left(|Y - X\theta| \leq t_2\right) = q_2$. Then we have

$$q_2 - q_1 = \mathbb{P}(t_1 < |Y - X\theta| \leq t_2) \geq \ell(t_2 - t_1).$$

Therefore, $\mathrm{F}_\theta^{-1}\left(q_2\right) - \mathrm{F}_\theta^{-1}\left(q_1\right) = t_2 - t_1 \leq \left(q_2 - q_1\right)/\ell$. □

If θ is the solution to β in the linear model, that is $Y - X\theta = \epsilon$. Thus the quantile function of ϵ satisfies 1-Hölder continuous.

Lemma 2. *Let* F_1 *and* F_2 *be two cumulative distribution functions on* $\mathbb{R}$*, and define*

$$\Delta := \sup_{t\in\mathbb{R}} \left|\mathrm{F}_1(t) - \mathrm{F}_2(t)\right|.$$

For a fixed $q \in (0,1)$ *such that* $[q - \Delta,\, q + \Delta] \subset (0,1)$*, suppose that the quantile functions* F_1^{-1} *and* F_2^{-1} *exist, and that* F_2^{-1} *is* γ*-Hölder continuous on* $[q-\Delta,\, q+\Delta]$ *with Hölder constant* $\mathcal{L}$ *for some* $\gamma \in (0,1]$*. Then it holds that*

$$\left|\mathrm{F}_1^{-1}(q) - \mathrm{F}_2^{-1}(q)\right| \;\leq\; \mathcal{L}\,\Delta^\gamma.$$

Proof. Fix $q \in (0,1)$ such that $[q - \Delta,\, q + \Delta] \subset (0,1)$. By definition of Δ,

for any $t \in \mathbb{R}$ we have

$$\left|\mathrm{F}_1(t) - \mathrm{F}_2(t)\right| \leq \Delta.$$

In particular, for $t = \mathrm{F}_1^{-1}(q)$, this implies $q - \Delta \leq \mathrm{F}_2\big(\mathrm{F}_1^{-1}(q)\big) \leq q + \Delta$. Hence,

$$\mathrm{F}_2^{-1}(q - \Delta) \leq \mathrm{F}_1^{-1}(q) \leq \mathrm{F}_2^{-1}(q + \Delta).$$

Subtracting $\mathrm{F}_2^{-1}(q)$ from all sides and using the γ-Hölder continuity of F_2^{-1} on $[q - \Delta,\, q + \Delta]$, we obtain

$$-\mathcal{L}\,\Delta^\gamma \leq\ \mathrm{F}_1^{-1}(q) - \mathrm{F}_2^{-1}(q) \leq \mathcal{L}\,\Delta^\gamma.$$

Therefore,

$$\left|\mathrm{F}_1^{-1}(q) - \mathrm{F}_2^{-1}(q)\right| \leq \mathcal{L}\,\Delta^\gamma.$$

□

Lemma 3. *Denote* F_n *as the empirical cumulative distribution function of* $|Y_i - X_i\theta|$*, and* $\mathrm{F}_{\widehat{n}}$ *as the empirical cumulative distribution function of* $|Y_i - X_i\widehat{\theta}|$*. On the event* $\{|X_i\widehat{\theta} - X_i\theta| \leq \rho_n\}$ *with* $\rho_n \geq 0$*, it holds*

$$|\mathrm{F}_n^{-1}(q) - \mathrm{F}_{\widehat{n}}^{-1}(q)| \leq \rho_n, \quad \forall q \in (0, 1).$$

Proof. On the event $\{|X_i\widehat{\theta} - X_i\theta| \leq \rho_n\}$, we have $\Big||Y_i - X_i\theta| - |Y_i - X_i\widehat{\theta}|\Big| \leq \rho_n$. According to the definition of the empirical cumulative distribution function, we have $\mathrm{F}_n(t) \leq \mathrm{F}_{\widehat{n}}(t + \rho_n)$. Let $t_1 = \mathrm{F}_n^{-1}(q)$ and $t_2 = \widehat{\mathrm{F}}_n^{-1}(q)$ be

the q-quantiles of F_n and $\mathrm{F}_{\widehat{n}}$ with $q \in (0,1)$ respectively. By the definition of the quantile function, for any $\varepsilon > 0$,

$$\mathrm{F}_n(t_1 - \varepsilon) < q \leq \mathrm{F}_n(t_1), \qquad \mathrm{F}_{\widehat{n}}(t_2 - \varepsilon) < q \leq \mathrm{F}_{\widehat{n}}(t_2).$$

Since $q \leq \mathrm{F}_n(t_1) \leq \mathrm{F}_{\widehat{n}}(t_1 + \rho_n)$, we have $t_2 \leq t_1 + \rho_n$. Similarly, we have $t_1 \leq t_2 + \rho_n$. Therefore, $|\mathrm{F}_n^{-1}(q) - \mathrm{F}_{\widehat{n}}^{-1}(q)| \leq \rho_n, \quad \forall q \in (0,1)$. □

Lemma 4. *Suppose random variables $\epsilon_1, \ldots, \epsilon_n$ are i.i.d., $\mathbb{E}\epsilon_1 = 0$, and there exists a constant $m > 0$ such that $\mathbb{E}\left|\epsilon_1\right|^m < \infty$. In addition, suppose the matrix $\Gamma = (\gamma_{ij})_{i=1,2,\ldots,k;\, j=1,2,\ldots,n}$ satisfies*

$$\max_{i=1,2,\ldots,k} \sum_{j=1}^{n} \gamma_{ij}^2 \leq D,$$

with $D > 0$. Then there exists a constant E_0, which only depends on m and $\mathbb{E}\left|\epsilon_1\right|^m$, such that for all $\tau > 0$,

$$\mathbb{P}\left(\max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ij}\epsilon_j \right| > \tau \right) \leq \frac{kE_0 D^{m/2}}{\tau^m}.$$

Proof. For any $i = 1, 2, \ldots, k$,

$$\mathbb{P}\left(\left| \sum_{j=1}^{n} \gamma_{ij}\epsilon_j \right| > \tau \right) \leq \frac{\mathbb{E}\left| \sum_{j=1}^{n} \gamma_{ij}\epsilon_j \right|^m}{\tau^m} \leq \frac{2^m C(m) \mathbb{E}\left|\epsilon_1\right|^m \left(\sum_{j=1}^{n} \gamma_{ij}^2 \right)^{m/2}}{\tau^m}$$
$$\leq \frac{2^m C(m) \mathbb{E}\left|\epsilon_1\right|^m D^{m/2}}{\tau^m}.$$

Choose $E_0 = 2^m C(m) \mathbb{E}\left|\epsilon_1\right|^m$, we have

$$\mathbb{P}\left(\max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ij}\epsilon_j \right| > \tau \right) \leq \sum_{i=1}^{k} \mathbb{P}\left(\left| \sum_{j=1}^{n} \gamma_{ij}\epsilon_j \right| > \tau \right) \leq \frac{kE_0 D^{m/2}}{\tau^m}.$$

□

S3 Proofs of theorems

Proof of Theorem 1. Let the prediction model $\widehat{\mu}$ and the bias correction term $\widehat{b}$ be estimated solely on the training set D_{train}. They are therefore fixed functions with respect to the calibration set D_{cal} and the test point (X_{n+1}, Y_{n+1}).

The nonconformity scores are generated via the transformation $T : (X_i, Y_i) \mapsto \widetilde{R}_i$. Since the underlying data are exchangeable, the resulting set of scores $\{\widetilde{R}_i\}_{i \in \mathcal{D}_{\text{cal}} \cup \{n+1\}}$ is also exchangeable. In other words, the rank of the test score is uniformly distributed on the set $\{1, 2, ..., n_{\text{cal}} + 1\}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{\text{DeCThRR}}(X_{n+1})\right) &= \mathbb{P}\left(\text{rank}(\widetilde{R}_{n+1}) \leq \lceil (1-\alpha)(n_{\text{cal}}+1) \rceil\right) \\
&= \frac{\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil}{n_{\text{cal}}+1} \\
&\geq 1 - \alpha.
\end{aligned}$$

To be easily understandable, we review the proof here as follows. Define $\alpha' = \alpha - 1/(n+1)$. By assuming a continuous joint distribution, and hence no ties among $\{\widetilde{R}_i\}_{i \in \mathcal{D}_{\text{cal}}}$, the set $\widehat{C}_{\text{DeCThRR}}\ (X_{n+1})$ excludes the values of y such that $|y - \widehat{\mu}\,(X_{n+1})|$ is among the $\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$ largest. Define

$\alpha' = \alpha - 1/(n_{\text{cal}}+1)$ and consider now the set $D(X_{n+1})$ consisting of points y such that $\widetilde{R}_i$ is among the $\lceil (n_{\text{cal}}+1)\alpha' \rceil$ largest. Then by construction

$$\mathbb{P}\left(Y_{n+1} \in D(X_{n+1})\right) \geq \alpha'.$$

$\widehat{C}_{\text{DeCThRR}}(X_{n+1}) \cap D(X_{n+1}) = \emptyset$, which implies the result. □

Proof of Theorem 2. We calculate

$$\begin{aligned}
\widehat{\theta} - \theta &= (\mathbf{X}^\top \mathbf{X} + h_n \mathbf{I}_p)^{-1} \boldsymbol{X}^\top y - \theta \\
&= \mathbf{Q}(\boldsymbol{\Lambda}^2 + h_n \mathbf{I}_r)^{-1} \boldsymbol{\Lambda} \mathbf{P}^\top (\mathbf{P} \boldsymbol{\Lambda} \mathbf{Q}^\top \beta + \boldsymbol{\epsilon}) - \mathbf{Q}\mathbf{Q}^\top \beta \qquad \text{(S3.1)} \\
&= -h_n \mathbf{Q}(\boldsymbol{\Lambda}^2 + h_n \mathbf{I}_r)^{-1} \zeta + \mathbf{Q}(\boldsymbol{\Lambda}^2 + h_n \mathbf{I}_r)^{-1} \boldsymbol{\Lambda} \mathbf{P}^\top \epsilon,
\end{aligned}$$

where $\zeta = \mathbf{Q}^\top \beta$. Denote $\widehat{\mathcal{M}}_{a_n} = \left\{ i \mid |\widehat{\theta}_i| > a_n \right\}$, we have

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{M}}_{a_n} \neq \mathcal{M}_{a_n}) &\leq \mathbb{P}(\min_{i \in \mathcal{M}_{a_n}} |\widehat{\theta}_i| \leq a_n) + \mathbb{P}(\max_{i \notin \mathcal{M}_{a_n}} |\widehat{\theta}_i| > a_n) \\
&\leq \mathbb{P}(\min_{i \in \mathcal{M}_{a_n}} |\theta_i| - \max_{i \in \mathcal{M}_{a_n}} |h_n \sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{\lambda_j^2 + h_n}| - \max_{i \in \mathcal{M}_{a_n}} |\sum_{j=1}^{r} \frac{q_{ij}\lambda_j}{\lambda_j^2 + h_n} \sum_{l=1}^{n} p_{lj}\epsilon_l| \leq a_n) \\
&+ \mathbb{P}(\max_{i \notin \mathcal{M}_{a_n}} |\theta_i| + \max_{i \notin \mathcal{M}_{a_n}} |h_n \sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{\lambda_j^2 + h_n}| + \max_{i \notin \mathcal{M}_{a_n}} |\sum_{j=1}^{r} \frac{q_{ij}\lambda_j}{\lambda_j^2 + h_n} \sum_{l=1}^{n} p_{lj}\epsilon_l| > a_n).
\end{aligned} \qquad \text{(S3.2)}$$

From Cauchy inequality,

$$\begin{aligned}
&\max_{i=1,\ldots,p} |h_n \sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{\lambda_j^2 + h_n}| \leq \max_{i=1,\ldots,p} h_n \sqrt{\sum_{j=1}^{r} q_{ij}^2} \sqrt{\sum_{j=1}^{r} \frac{\zeta_j^2}{(\lambda_j^2 + h_n)^2}} = O(n^{\alpha_\theta - 2\eta + \delta}), \\
&\max_{i=1,\ldots,p} \sum_{l=1}^{n} (\sum_{j=1}^{r} \frac{q_{ij}\lambda_j}{\lambda_j^2 + h_n} p_{lj})^2 = \max_{i=1,\ldots,p} \sum_{j=1}^{r} \frac{q_{ij}^2 \lambda_j^2}{(\lambda_j^2 + h_n)^2} \leq \max_{i=1,\ldots,p} \frac{\sum_{j=1}^{r} q_{ij}^2}{\lambda_r^2}.
\end{aligned} \qquad \text{(S3.3)}$$

Therefore, for sufficiently large n, from assumptions B5 and B6, we have

$$\begin{aligned}
&\min_{i\in\mathcal{M}_{a_n}} |\widehat{\theta}_i| - \max_{i\in\mathcal{M}_{a_n}} |h_n \sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{\lambda_j^2+h_n}| - a_n \geq \frac{1}{2}(\frac{1}{c_a}-1)a_n, \\
&a_n - \max_{i\notin\mathcal{M}_{a_n}} |\widehat{\theta}_i| - \max_{i\notin\mathcal{M}_{a_n}} |h_n \sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{\lambda_j^2+h_n}| < \frac{1}{2}(1-c_a)a_n.
\end{aligned} \tag{S3.4}$$

From Lemma 4, there exist constants E_1 and E_2 depending on m such that

$$\mathbb{P}(\widehat{\mathcal{M}}_{a_n} \neq \mathcal{M}_{a_n}) \leq \frac{|\mathcal{M}_{a_n}| \times E_1}{\lambda_r^m \times (\frac{1}{2}(\frac{1}{c_a}-1)a_n)^m} + \frac{(p-|\mathcal{M}_{a_n}|) \times E_2}{\lambda_r^m \times (\frac{1}{2}(1-c_a)a_n)^m} = O(n^{\alpha_p - m\eta + m\alpha_a}). \tag{S3.5}$$

The consistency of variable selection is verified by Assumption B5. If $\widehat{\mathcal{M}}_{a_n} = \mathcal{M}_{a_n}$,

$$|X_i\widehat{\theta} - X_i\theta| \leq |h_n \sum_{j=1}^{r} c_{ij}\frac{1}{\lambda_j^2+h_n}\zeta_j| + |\sum_{j=1}^{r} c_{ij}\frac{\lambda_j}{\lambda_j^2+h_n}\sum_{l=1}^{n} p_{lj}\epsilon_l| \tag{S3.6}$$

From Cauchy inequality,

$$|h_n \sum_{j=1}^{r} c_{ij}\frac{1}{\lambda_j^2+h_n}\zeta_j| \leq h_n \sqrt{\sum_{j=1}^{r} c_{ij}^2}\sqrt{\sum_{j=1}^{r}(\frac{1}{\lambda_j^2+h_n}\zeta_j)^2} = O(n^{\alpha_\theta - 2\eta + \delta}). \tag{S3.7}$$

Since $\sum_{l=1}^{n}(\sum_{j=1}^{r} c_{ij}\frac{\lambda_j}{\lambda_j^2+h_n}\sum_{l=1}^{n} p_{lj})^2 = \sum_{j=1}^{r} c_{ij}^2\frac{\lambda_j^2}{(\lambda_j^2+h_n)^2} \leq \frac{C_{\alpha_\mathcal{N}}}{\lambda_r^2}$, we have

$$\begin{aligned}
\mathbb{P}(|\sum_{j=1}^{r} c_{ij}\frac{\lambda_j}{\lambda_j^2+h_n}\sum_{l=1}^{n} p_{lj}\epsilon_l| > \tau) &\leq \frac{\mathbb{E}(|\sum_{j=1}^{r} c_{ij}\frac{\lambda_j}{\lambda_j^2+h_n}\sum_{l=1}^{n} p_{lj}\epsilon_l|^m)}{\tau^m} \\
&\leq \frac{2^m C(m)\mathbb{E}|\epsilon_1|^m \left(\sum_{j=1}^{r}(c_{ij}\frac{\lambda_j}{(\lambda_j^2+h_n}\sum_{l=1}^{n} p_{lj})^2\right)^{m/2}}{\tau^m} \\
&\leq \frac{2^m C(m)\mathbb{E}|\epsilon_1|^m C_{\alpha_\mathcal{N}}^{m/2}}{\tau^m \lambda_r^m}.
\end{aligned} \tag{S3.8}$$

Choose a constant $C = 2^m C(m)\mathbb{E}|\epsilon_1|^m$, we have $|\sum_{j=1}^{r} c_{ij}\frac{\lambda_j}{\lambda_j^2+h_n}\sum_{l=1}^{n} p_{lj}\epsilon_l| = O_p(n^{-\eta})$. Therefore, we prove $|X_i\widehat{\theta}-X_i\theta| = O_p(n^{\alpha_\theta-2\eta+\delta}+n^{-\eta}) = O_p(n^{\alpha_\theta-2\eta+\delta})$ according to Assumption B5. Since $|Y_i - X_i\theta| - |Y_i - X_i\widehat{\theta}| \leq |X_i\widehat{\theta} - X_i\theta|$, for $\forall\epsilon \in (0,1)$, there exist a constant $c_3 > 0$ such that

$$|Y_i - X_i\widehat{\theta}| - |Y_i - X_i\theta| \leq c_3 n^{\alpha_\theta-2\eta+\delta}, \quad i = 1, ..., n, \tag{S3.9}$$

with at least $1-\epsilon$ probability.

Denote F_n as empirical cumulative distribution function of $|Y_i - X_i\theta|$, and F_0 is the distribution function of $|Y_i - X_i\theta|$. $\mathrm{F}_{\widehat{n}}$ is empirical CDF of $|Y_i - X_i\widehat{\theta}|$ and F_1 is the distribution function of $|Y_i - X_i\widehat{\theta}|$. In the following proof, we will achieve our conclusion through three main steps. First we clarify the relationship between F_0^{-1} and F_1^{-1}. Next use DKW Theorem bound the discrepancy between the inverse of the empirical distribution and the inverse of the true distribution, then analyze the relationship between the two empirical distributions. Finally combine the results of the previous steps to conclude the proof.

On the event $|Y_i - X_i\widehat{\theta}| - |Y_i - X_i\theta| \leq c_3 n^{\alpha_\theta-2\eta+\delta}$, by Assumption A3,

the density of $|Y_i - X_i\theta|$ is bounded by M, which implies that

$$\begin{aligned}|\mathrm{F}_1(t) - \mathrm{F}_0(t)| &= \left|\mathbb{P}(|Y_i - X_i\widehat{\theta}| < t) - \mathbb{P}(|Y_i - X_i\theta| < t)\right| \\ &\leq \left|\mathbb{P}(|Y_i - X_i\theta| - c_3 n^{\alpha_\theta - 2\eta + \delta} < t) - \mathbb{P}(|Y_i - X_i\theta| < t)\right| \\ &\leq M c_3 n^{\alpha_\theta - 2\eta + \delta}.\end{aligned} \tag{S3.10}$$

The first inequality follows from (S3.9).

Since the density of $|Y_i - X_i\theta|$ is bounded, the inverse of F_0 is 1-Hölder continuous by Lemma 1. Therefore, we can use (S3.10) and Lemma 2 to obtain

$$|\mathrm{F}_1^{-1}(1-\alpha) - \mathrm{F}_0^{-1}(1-\alpha)| \leq c_4 n^{\alpha_\theta - 2\eta + \delta}, \tag{S3.11}$$

where $c_4 > 0$ is a constant. Applying the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality and Lemma 2, we have

$$|\mathrm{F}_{\widehat{n}}^{-1}(1-\alpha) - \mathrm{F}_1^{-1}(1-\alpha)| = O_p(n^{-1/2}). \tag{S3.12}$$

Therefore, combining with above inequalities (S3.10) and (S3.12), we have $|\mathrm{F}_{\widehat{n}}^{-1}(1-\alpha) - q_{1-\alpha}| = |\mathrm{F}_{\widehat{n}}^{-1}(1-\alpha) - \mathrm{F}_0^{-1}(1-\alpha)| = O_p(n^{\alpha_\theta - 2\eta + \delta})$. □

Proof of Theorem 3. We first introduce the debiased estimator, denoted by $\widetilde{\theta}^* = \widetilde{\theta} + h_n X_i \boldsymbol{Q}(\boldsymbol{\Lambda}^2 + h_n \mathbf{I}_r)^{-1}\mathbf{Q}^\top\widetilde{\theta}$. This estimator is constructed to reduce the bias inherent in the standard ridge estimator $\widehat{\theta}$. The estimation error

of this debiased estimator is given by

$$\begin{aligned}\widetilde{\theta}^* - \theta = &-h_n^2 \mathbf{Q}(\mathbf{\Lambda}^2 + h_n \mathbf{I}_r)^{-2}\zeta \\ &+ \mathbf{Q}\left((\mathbf{\Lambda}^2 + h_n \mathbf{I}_r)^{-1}\mathbf{\Lambda} + h_n(\mathbf{\Lambda}^2 + h_n \mathbf{I}_r)^{-2}\mathbf{\Lambda}\right)\mathbf{P}^\top \epsilon.\end{aligned} \tag{S3.13}$$

Following an analysis similar to that in the proof of Theorem 2, which leverages the assumptions on the parameters, we can show that the prediction error of the debiased estimator converges at a faster rate as

$$|X_i\widetilde{\theta}^* - X_i\theta| = O_p(n^{\alpha_\theta - 4\eta + 2\delta} + n^{-\eta}) = O_p(n^{-\eta}), \tag{S3.14}$$

where the final equality holds under Assumption B5, which implies $\alpha_\theta - 4\eta + 2\delta < -\eta$.

The debiased conformal prediction band is then constructed using this improved estimator. The nonconformity scores are recalculated as $\widetilde{R}_i = |Y_i - X_i\widetilde{\theta}^*|$, and the prediction interval for a new point X_{n+1} is given by

$$\widehat{C}_{\text{DeCThRR}}(X_{n+1}) = \left[X_{n+1}\widetilde{\theta}^* \pm Q_{1-\alpha}\left(\{\widetilde{R}_i\}_{i\in\mathcal{D}_{\text{cal}}}\right)\right],$$

where $Q_{1-\alpha}$ denotes the appropriate empirical quantile (e.g., the $\lceil(1-\alpha)(n_{\text{cal}}+1)\rceil$-th smallest value) of the calibration scores. The remainder of the proof follows the same structure as the proof of Theorem 2, but starts from the faster convergence rate in (S3.14). □

Proof of Theorem 4. The proof that $L(\widehat{C}_{\text{DeCThRR}}(X_{n+1})\Delta C_o^*(x)) = o_p(1)$ consists of showing that the centers and widths of the two intervals converge.

Step 1: Asymptotic Convergence of Interval Centers. The center of our proposed prediction interval is $C_{\text{center}} = X_{n+1}\tilde{\theta}^*$, while the center of the oracle interval is $C^*_{o,\text{center}} = X_{n+1}\theta$. From the analysis in the proof of Theorem 3, specifically equation (S3.14), we have established the convergence rate of the prediction error as

$$|C_{\text{center}} - C^*_{o,\text{center}}| = |X_{n+1}(\tilde{\theta}^* - \theta)| = O_p(n^{-\eta}).$$

Since $\eta > 0$, the difference between the centers converges to zero in probability, i.e., $|C_{\text{center}} - C^*_{o,\text{center}}| = o_p(1)$.

Step 2: Asymptotic Equivalence of Interval Lengths. The width of our interval is $\text{Width}(\widehat{C}_{\text{DeCThRR}}) = 2\text{F}^{-1}_{\tilde{n}}(1-\alpha)$, where $\text{F}^{-1}_{\tilde{n}}(1-\alpha)$ is the empirical quantile of the debiased residuals. The width of the oracle interval is $\text{Width}(C^*_o) = 2q_{1-\alpha}$. Theorem 3 directly establishes the convergence of the interval width as

$$\text{Width}(\widehat{C}_{\text{DeCThRR}}(X_{n+1})) - \text{Width}(C^*_o(x)) = O_p(n^{-\eta}).$$

This implies that the difference between the lengths also converges to zero in probability. □

S4 Experiments Results

This appendix provides supplementary figures complementing the simulation study described in Section 4. We also conduct Case 1S, a coordinate-wise sparse variant of Case 1, keeping the same design and noise but changing $\boldsymbol{\beta}$. This scenario represents the classical coordinate-wise sparse regime, which is typically favorable to methods such as Lasso that assume sparsity at the individual-coordinate level, while still maintaining strong predictor correlations to challenge variable-selection stability. The data are generated from the linear model $Y = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0,\, 1.5^2\mathbf{I}_n)$. The design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ is sampled from a zero-mean multivariate normal distribution with the same block-wise covariance structure as in Case 1: predictors are partitioned into blocks of size 20 with strong intra-block correlation (0.4), and a mild cross-block dependence is induced by setting correlation 0.15 between each predictor and the first five predictors in the subsequent block. The regression coefficient vector $\boldsymbol{\beta}$ is coordinate-wise sparse: we uniformly sample a support set $S \subset \{1, \ldots, p\}$ with $|S| = s = 20$ and set

$$\beta_j = 1.5\,\xi_j \cdot \mathbf{1}(j \in S), \qquad \xi_j \in \{-1, +1\} \text{ i.i.d..}$$

This case evaluates how different conformal pipelines behave in the standard sparse setting under correlated covariates, and in particular whether

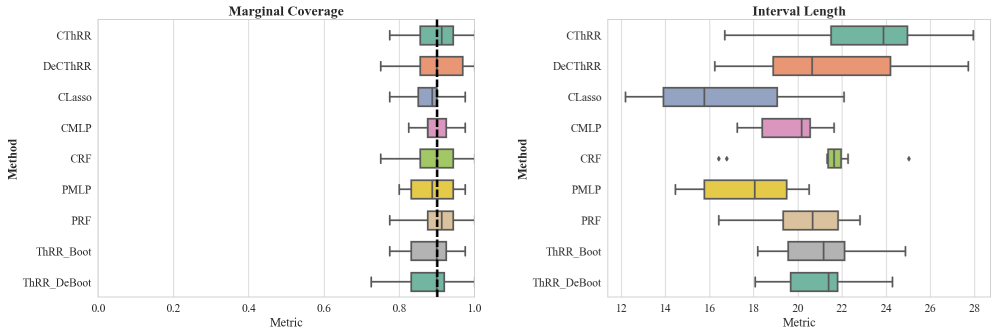
debiasing within the conformal score construction can retain calibration and produce informative intervals when the signal is coordinate-wise sparse.

As outlined in the main text, we consider five distinct data-generating mechanisms that span sparse, dense, nonlinear, and highly correlated regimes. All hyperparameters are tuned via 5-fold cross-validation on the training set, the nominal coverage level is $1 - \alpha = 0.90$, and each experiment is repeated over 200 independent replications.

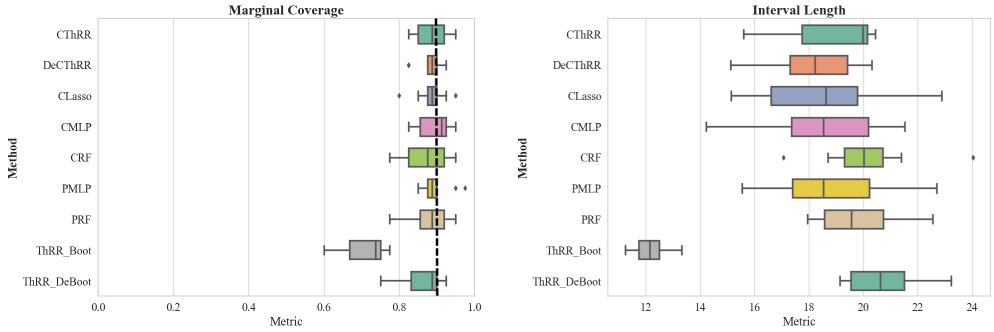
While Section 4 presents detailed results for Case 1, here we report the corresponding boxplots for Case 1S and Cases 2–4. We also examine the width of the prediction intervals as the ridge parameter h_n varies, for settings with $p = 500$ and $p = 1000$ in Cases 2–4. These supplementary results reinforce our main findings in Section 4: the proposed DeCThRR method consistently produces shorter intervals and achieves improved coverage compared to CThRR, particularly in dense-signal and multicollinear settings (Cases 2 and 4), while also delivering robust inference under model misspecification (Case 3). Moreover, the bias-correction step further contributes to narrowing the prediction intervals, as illustrated in Figure S5.

Overall, these supplementary results further confirm the performance of our proposed DeCThRR method across diverse data-generating mechanisms. In particular, they demonstrate its systematic advantage over

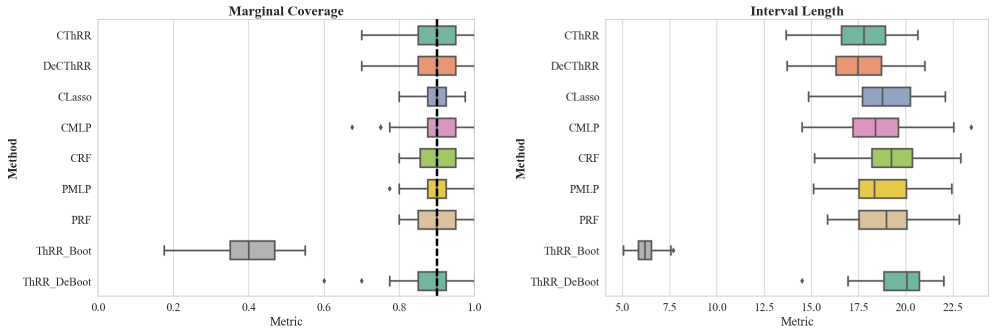
CThRR in terms of achieving narrower prediction intervals with reliable coverage, especially in dense, multicollinear, and even in misspecified settings.



(a) Coverage and interval width when $p = 500, n = 200$ in Case 1S.

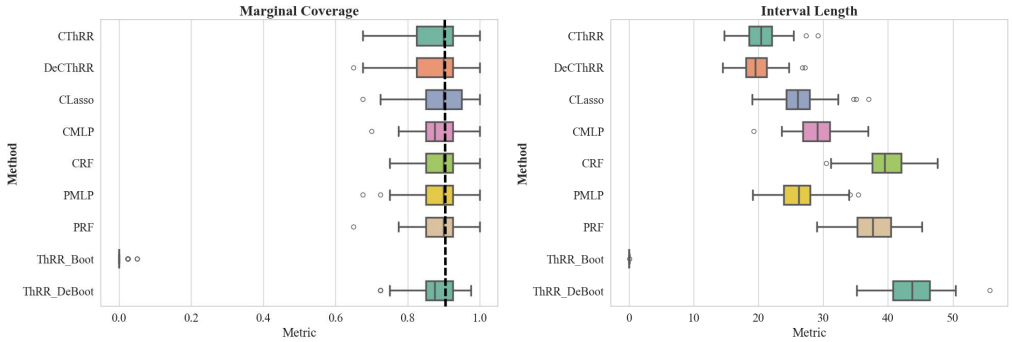


(b) Coverage and interval width when $p = 1000, n = 200$ in Case 1S.

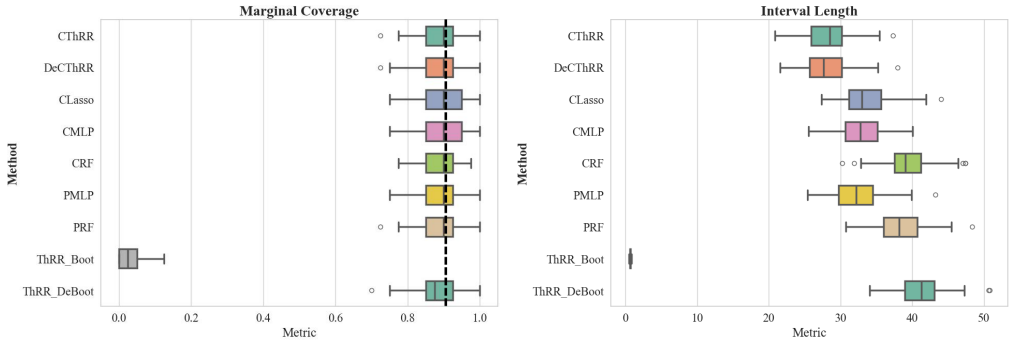


(c) Coverage and interval width when $p = 1500, n = 200$ in Case 1S.

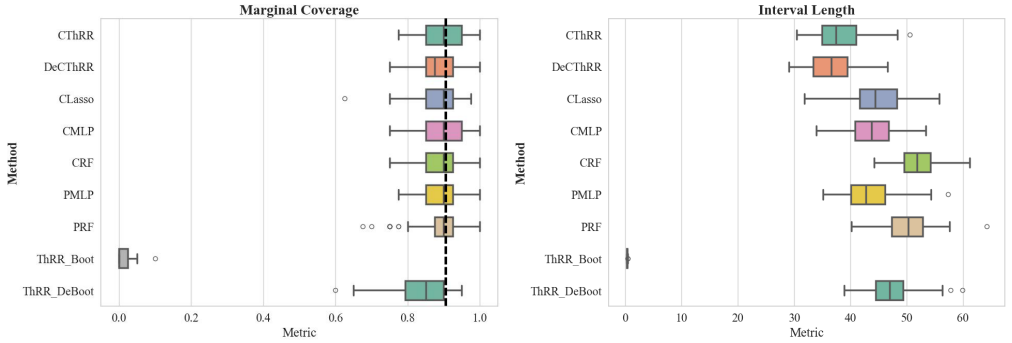
Figure S1: Average coverage and interval length of methods in Case 1S.



(a) Coverage and interval width when $p = 500, n = 200$ in Case 2.

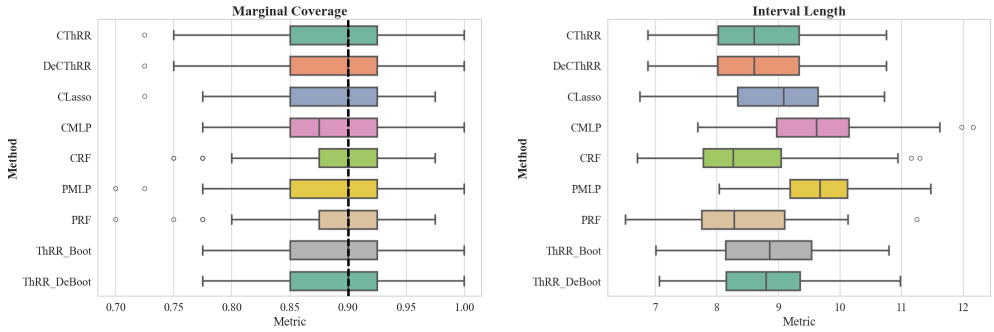


(b) Coverage and interval width when $p = 1000, n = 200$ in Case 2.

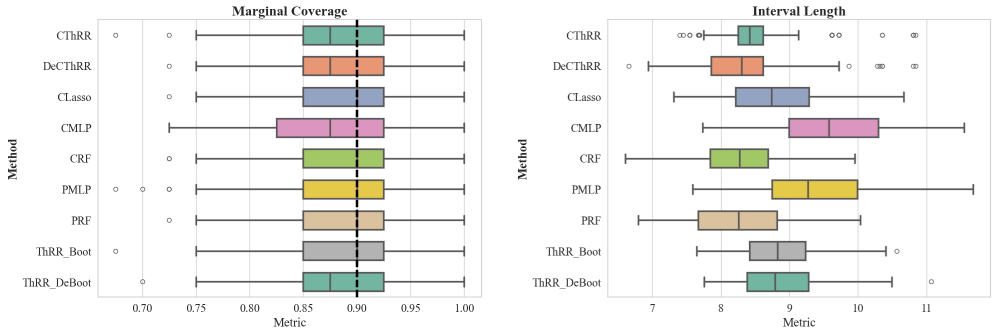


(c) Coverage and interval width when $p = 1500, n = 200$ in Case 2.

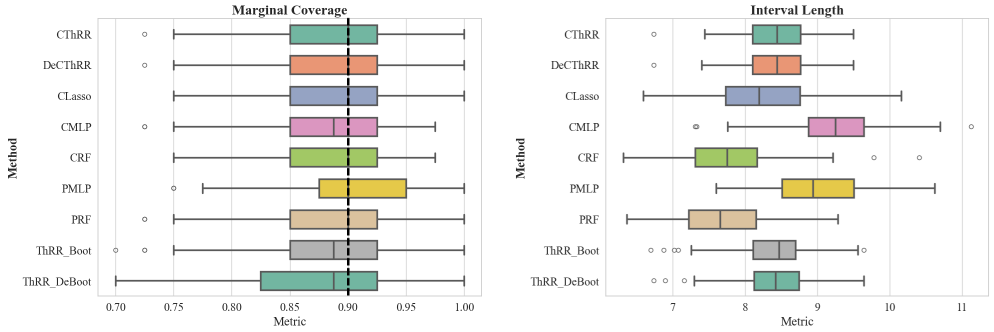
Figure S2: Average coverage and interval length of methods in Case 2.



(a) Coverage and interval width when $p = 500, n = 200$ in Case 3.

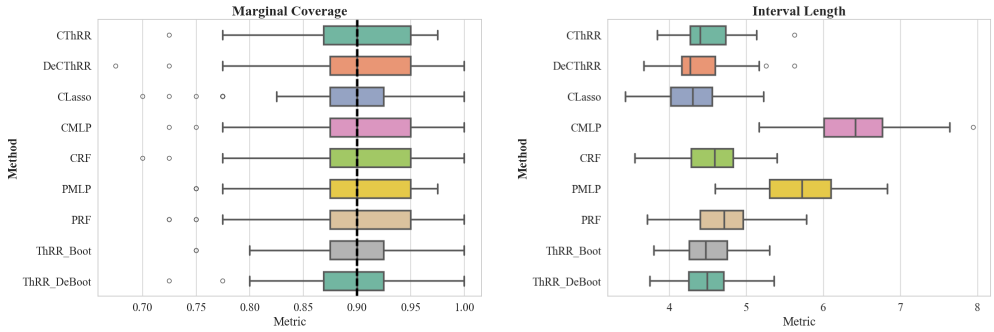


(b) Coverage and interval width when $p = 1000, n = 200$ in Case 3.

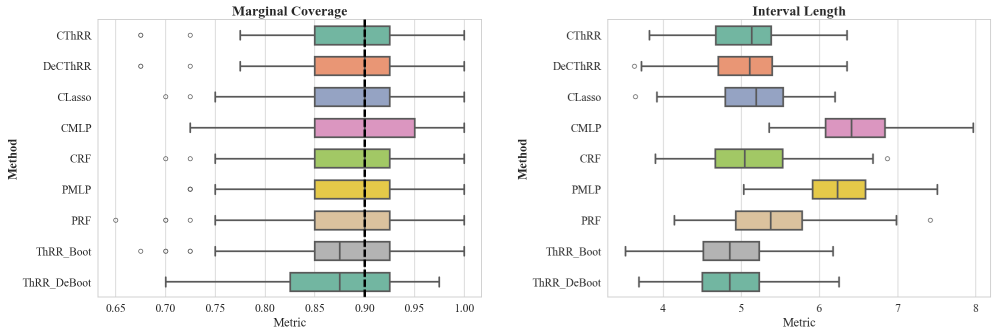


(c) Coverage and interval width when $p = 1500, n = 200$ in Case 3.

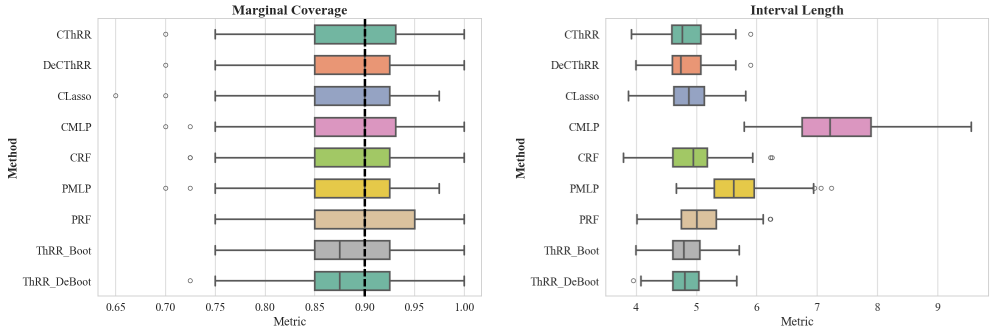
Figure S3: Average Coverage and interval length of methods in Case 3.



(a) Coverage and interval width when $p = 500, n = 200$ in Case 4.

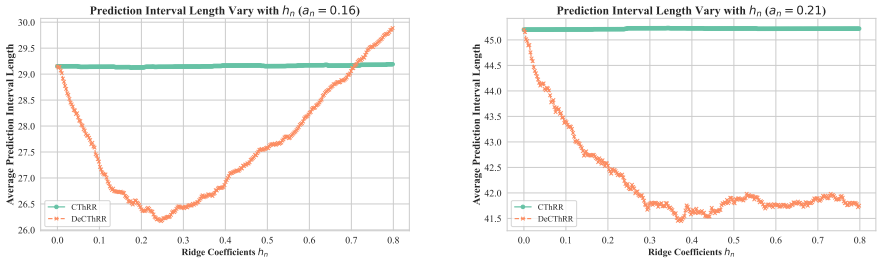


(b) Coverage and interval width when $p = 1000, n = 200$ in Case 4.

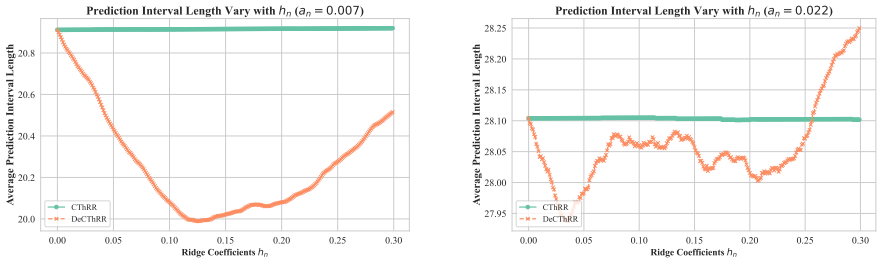


(c) Coverage and interval width when $p = 1500, n = 200$ in Case 4.

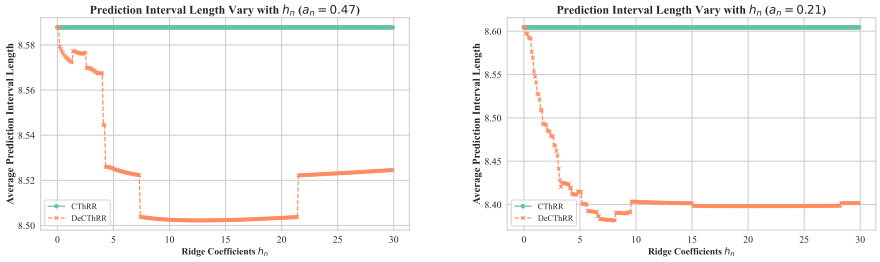
Figure S4: Average Coverage and interval length of methods in Case 4.



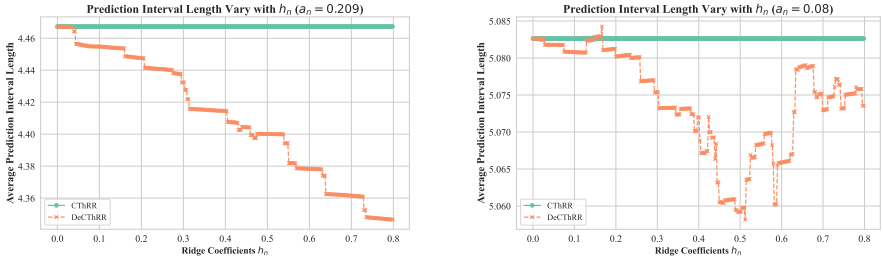
(a) Case 1.



(b) Case 2.



(c) Case 3.



(d) Case 4.

Figure S5: Prediction interval length of ThRR within debiasing conformal prediction and original counterpart over the different cases. The left figures are Cases when $p = 500$ and the right figures are Cases when $p = 1000$.

Table S1: Average interval length, empirical coverage (in parentheses) and average runtime across $p = 500, 1000, 1500$ for Case 1S.

	$p = 500$	$p = 1000$	$p = 1500$	Time (s)
CThRR	18.921(0.888)	22.931(0.903)	17.711(0.901)	0.0125
DeCThRR	**18.107(0.883)**	21.380(0.905)	**17.498(0.895)**	0.1321
CLasso	18.426(0.885)	**16.428(0.875)**	18.774(0.898)	**0.0115**
CMLP	18.337(0.895)	19.618(0.893)	18.477(0.894)	0.0900
CRF	20.154(0.868)	21.043(0.898)	19.264(0.908)	1.6167
PMLP	18.807(0.895)	17.730(0.888)	18.661(0.903)	0.4779
PRF	19.813(0.875)	20.394(0.908)	18.956(0.899)	0.3242
ThRR_Boot	12.197(0.710)	21.124(0.885)	6.201(0.407)	4.4097
ThRR_DeBoot	20.689(0.863)	21.089(0.875)	19.671(0.887)	24.8640

Table S2: Average interval length, empirical coverage (in parentheses) and average runtime across $p = 500, 1000, 1500$ for Case 2.

	$p = 500$	$p = 1000$	$p = 1500$	Time (s)
CThRR	20.310(0.877)	28.289(0.889)	38.152(0.893)	**0.018**
DeCThRR	**19.613(0.876)**	**27.997(0.890)**	**36.657(0.885)**	0.163
CLasso	26.371(0.889)	33.706(0.895)	44.377(0.889)	0.217
CMLP	29.033(0.882)	32.781(0.889)	43.790(0.896)	0.539
CRF	39.674(0.883)	39.415(0.889)	51.917(0.890)	1.471
PMLP	26.284(0.882)	32.384(0.887)	43.421(0.892)	0.520
PRF	37.742(0.885)	38.421(0.883)	50.175(0.888)	0.369
ThRR_Boot	0.067(0.006)	0.675(0.025)	0.300(0.010)	2.677
ThRR_DeBoot	43.349(0.870)	41.140(0.878)	47.200(0.836)	41.152

Table S3: Average interval length, empirical coverage (in parentheses) and average runtime across $p = 500, 1000, 1500$ for Case 3.

	$p = 500$	$p = 1000$	$p = 1500$	Time (s)
CThRR	8.725(0.886)	8.612(0.877)	8.458(0.889)	**0.015**
DeCThRR	8.687(0.888)	8.403(0.876)	8.362(0.889)	0.163
CLasso	9.000(0.887)	8.740(0.888)	8.278(0.892)	0.062
CMLP	9.687(0.886)	9.626(0.878)	9.221(0.882)	0.258
CRF	8.453(0.892)	8.331(0.885)	7.797(0.889)	1.424
PMLP	9.653(0.885)	9.392(0.886)	9.002(0.897)	0.172
PRF	**8.373(0.893)**	**8.240(0.886)**	**7.728(0.888)**	0.236
ThRR_Boot	8.907(0.889)	8.859(0.883)	8.375(0.880)	2.403
ThRR_DeBoot	8.887(0.887)	8.870(0.883)	8.383(0.876)	22.760

Table S4: Average interval length, empirical coverage (in parentheses) and average runtime across $p = 500, 1000, 1500$ for Case 4.

	$p = 500$	$p = 1000$	$p = 1500$	Time (s)
CThRR	4.507(0.898)	5.079(0.885)	4.846(0.892)	**0.016**
DeCThRR	**4.373(0.897)**	5.007(0.885)	**4.795(0.897)**	0.157
CLasso	4.411(0.895)	5.188(0.880)	4.866(0.887)	0.050
CMLP	6.425(0.897)	6.448(0.889)	7.290(0.885)	0.291
CRF	4.568(0.896)	5.123(0.885)	4.931(0.891)	1.553
PMLP	5.712(0.897)	6.172(0.887)	5.683(0.887)	0.068
PRF	4.706(0.898)	5.384(0.887)	5.031(0.891)	0.206
ThRR_Boot	4.505(0.892)	4.975(0.877)	4.819(0.882)	2.055
ThRR_DeBoot	4.503(0.895)	**4.965(0.876)**	4.801(0.881)	26.674

Bibliography

Bai, Z. and Y. Yin (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability 21*(3), 1275–1294.

Shao, J. and X. Deng (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics 40*(2), 812–831.

Zhang, Y. and D. N. Politis (2022). Ridge regression revisited: Debiasing, thresholding and bootstrap. *The Annals of Statistics 50*(3), 1401–1422.