Lei et al. (2018)

Qianyu Dong, Toshiya Yoshida

University of California, Santa Cruz

January 18, 2024

Paper of the week

- Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman (2018). Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association, 113(523), 1094-1111.

Agenda

Introduction

- Consider i.i.d. regression data $Z_1, ..., Z_n \sim P$, where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$, response variable Y_i and covariates $X_i = (X_i(1), ..., X_i(d))$. Let $\mu(x) = \mathbb{E}(Y|X = x),\ x \in \mathbb{R}^d$ denote the regression function
- We are interested in predicting a new response Y_{n+1} from a new feature value X_{n+1}.
- Formally: prediction band $C \subseteq \mathbb{R}^d \times \mathbb{R}$, for given miscoverage level $\alpha \in (0, 1)$

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha, \tag{1}$$

without assumptions on $\mu(x)$ and P

Conformal prediction

- The conformal prediction framework was originally proposed as a sequential approach for forming prediction intervals by Vovk et al. (2005).
- Testing the null hypothesis that $Y_{n+1} = y$ and construct a valid p-value based on the empirical quantiles of the augmented sample $(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ with $Y_{n+1} = y$
- Proper finite-sample coverage without any assumptions on P and $\hat{\mu}$, except that $\hat{\mu}$ act a symmetric function and exchangeability of data points.

Conformal prediction

- For each value $y \in \mathbb{R}$, construct an augmented regression estimator $\hat{\mu}_y$, which is trained on the augmented data set $Z_1, ..., Z_n, (X_{n+1}, y)$
- Define

$$R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|, \ \ i = 1, ..., n$$

$$R_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})| \tag{2}$$

- Rank $R_{y,n+1}, R_{y,1}, \cdots R_{y,n}$ and compute

$$\pi(y) = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}\{R_{y,i} \leq R_{y,n+1}\} \tag{3}$$

  - The proportion of points in the augmented sample whose fitted residual is smaller than $R_{y,n+1}$
  - $\pi(Y_{n+1})$ is uniformly distributed over the set $(1/(n+1), 2/(n+1), \cdots, 1)$, because of exchangeability of the data points and symmetry of $\hat{\mu}$.

Conformal prediction cont.

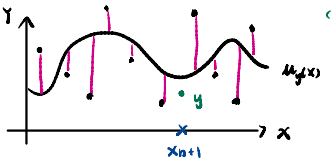
- $(n+1)\pi(Y_{n+1})$ is uniformly distributed over the set $(1, 2, \cdots, (n+1))$ implies

$$P\left((n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil\right) \geq 1-\alpha \tag{4}$$

  - We may interpret the $1-\pi(Y_{n+1})$ provides a valid (conservative) p-value for testing the null hypothesis that $H0 : Y_{n+1} = y$.
- By inverting the test, conformal prediction interval at X_{n+1} is

$$C_{\text{conf}}(X_{n+1}) = \{y \in \mathbb{R} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\} \tag{5}$$

- Assumptions:
  - $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable
  - Model $\mu(x)$ is symmetric. Order of the data does not matter.



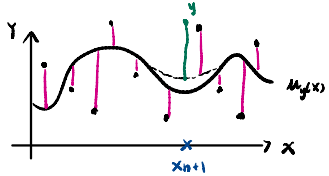
(n = 10)

$$R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|,\ i = 1, ..., n$$

$$R_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})|$$

$\pi(y) = \frac{1+3}{1+n}$

$$\pi(y) = \frac{1}{n+1} + \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{R_{y,i} \le R_{y,n+1}\}$$



$\pi(y) = \frac{1+9}{1+n}$

b/c ex. $\pi(y) \sim$ unif on $\{\frac{1}{1+n}, \ldots \frac{1+n}{1+n}\}$

$(n-1)\pi(y) \sim$ unif $\{1, 2, \ldots n\}$

$P(L(n-1)\pi(y)\rfloor \le \lceil (n-1)(1-\alpha)\rceil) \le 1-\alpha$

$C_{conf} = \{y \in \mathbb{R} : (n+1)\pi(y) \le \lceil (1-\alpha)(n-1)\rceil\}$

Algorithm 1

Algorithm 1 Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$, points $\mathcal{X}_{\text{new}} = \{X_{n+1}, X_{n+2}, \ldots\}$ at which to construct prediction intervals, and values $\mathcal{Y}_{\text{trial}} = \{y_1, y_2, \ldots\}$ to act as trial values

Output: Predictions intervals, at each element of $\mathcal{X}_{\text{new}}$

for $x \in \mathcal{X}_{\text{new}}$ **do**

  for $y \in \mathcal{Y}_{\text{trial}}$ **do**

    $\widehat{\mu}_y = \mathcal{A}\big(\{(X_1, Y_1), \ldots, (X_n, Y_n), (x, y)\}\big)$

    $R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|$, $i = 1, \ldots, n$, and $R_{y,n+1} = |y - \widehat{\mu}_y(x)|$

    $\pi(y) = (1 + \sum_{i=1}^{n} \mathbb{1}\{R_{y,i} \leq R_{y,n+1})\}/(n+1)$

  end for

  $C_{\text{conf}}(x) = \{y \in \mathcal{Y}_{\text{trial}} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\}$

end for

Return $C_{\text{conf}}(x)$, for each $x \in \mathcal{X}_{\text{new}}$

Theorem 1

Theorem

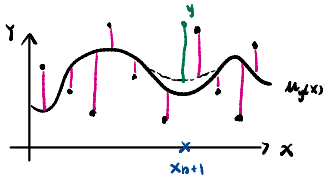
If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. pair (X_{n+1}, Y_{n+1}), the lower bound holds for the conformal prediction band C_{conf} constructed in (5). If we assume additionally that for all $y \in \mathbb{R}$, the fitted absolute residuals $R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|$, $i = 1, \ldots, n$ have a continuous joint distribution, then the upper holds that

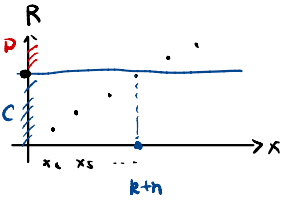
$$1 - \alpha \leq \mathbb{P}(Y_{n+1} \in C_{conf}(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}.$$

Proof.

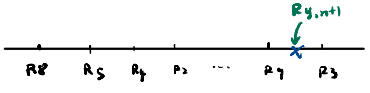
The lower bound holds by the the construction of the conformal prediction set. For upper bound, constrct $D(X_{n+1})$ consisting of points y that are excluded in $C_{\text{conf}}(X_{n+1})$

□





Pf: Assume a continuous joint distribution for fitted residuals $\Rightarrow P(R_{y,i} = R_{y,j}) = 0 \Rightarrow$ distinct $R_{y,1}, \dots R_{y,n}$



$\lceil (n+1)(1-\alpha) \rceil$ smallest

$\lceil (n+1)\alpha - 1 \rceil$ largest

$n+1 \geq (1-\alpha+\alpha) - 1$

consider $\alpha' = \alpha - \frac{1}{n+1}$

consider $D(X_{n+1})$ consisting y s.t. $R_{y,n+1}$ is $\lceil (n+1)\alpha' \rceil$ largest

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

$C \cap D = \emptyset$

$P(Y_{n+1} \in C(X_{n+1})) \leq P(Y_{n+1} \notin D(X_{n+1}))$

$= 1 - P(Y_{n+1} \in D(X_{n+1}))$

$\leq 1 - \alpha'$

$= 1 - \alpha + \frac{1}{n+1}$ □

Remarks

- The first part of the theorem, is a standard property of all conformal inference procedures and is due to Vovk. The second part is new. Other than the continuity assumption, no assumptions are needed of $\hat{\mu}$ and P.
- Data are not exchangeable, lose uniform distribution properties of $\pi(y)$. Tibshirani et al. (2019)
- Improve underlying regression function μ, the resulting conformal prediction interval decreases in length.
- marginal coverage guarantees.
- *Theorem 1* holds for replacing each $R_{y,i}$ by

$$f((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_{n+1}, y); (X_i, Y_i)),$$

where f is any function that is symmetric in its first n arguments. It is conformity score in the context of conformal inference.

Split Conformal Prediction Sets

- $$R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|, \ i = 1, ..., n$$
$$R_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})| \quad (2)$$

  The full conformal prediction method computationally intensive. We need retrain the model on the augmented data set for each y.
- In some applications, X_{n+1} is not necessarily observed.
- Consider **Split Conformal Prediction**, which separates the fitting and ranking steps using sample splitting.
- Similar ideas: inductive conformal inference (Papadopoulos et al., 2002; Vovk et al., 2005)

Algorithm 2

Algorithm 2 Split Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1$, $\mathcal{I}_2$
$\widehat{\mu} = \mathcal{A}\big(\{(X_i, Y_i) : i \in \mathcal{I}_1\}\big)$
$R_i = |Y_i - \widehat{\mu}(X_i)|$, $i \in \mathcal{I}_2$
$d =$ the kth smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where $k = \lceil (n/2 + 1)(1 - \alpha) \rceil$
Return $C_{\text{split}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

Theorem 2

Theorem (Theorem 2)

If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. draw (X_{n+1}, Y_{n+1}),

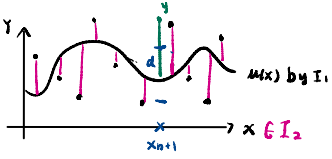
$$P(Y_{n+1} \in C_{split}(X_{n+1})) \geq 1 - \alpha,$$

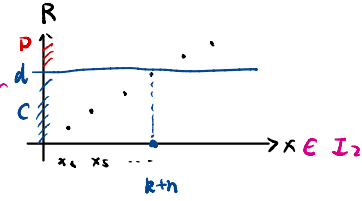
for the split conformal prediction band C_{split} constructed in Algorithm 2. Moreover, if we assume additionally that the residuals R_i, $i \in I_2$ have a continuous joint distribution, then

$$P(Y_{n+1} \in C_{split}(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2}.$$

Algorithm 2 Split Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1, \mathcal{I}_2$
$\widehat{\mu} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$
$R_i = |Y_i - \widehat{\mu}(X_i)|$, $i \in \mathcal{I}_2$
$d =$ the kth smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where $k = \lceil (n/2+1)(1-\alpha) \rceil$
Return $C_{\text{split}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$



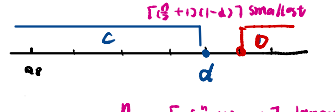


Theorem 2.2. *If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. draw (X_{n+1}, Y_{n+1}),*

$$\mathbb{P}(Y_{n+1} \in C_{\text{split}}(X_{n+1})) \geq 1 - \alpha,$$

for the split conformal prediction band C_{split} constructed in Algorithm 2. Moreover, if we assume additionally that the residuals R_i, $i \in \mathcal{I}_2$ have a continuous joint distribution, then

$$\mathbb{P}(Y_{n+1} \in C_{\text{split}}(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2}.$$



$\frac{n}{2} - \lceil (\frac{n}{2}+1)(1-\alpha) \rceil$ largest

$P(Y_{n+1} \in D(X_{n+1}))$

$\geq \dfrac{\frac{n}{2} - \lceil (\frac{n}{2}+1)(1-\alpha) \rceil}{\frac{n}{2}+1}$

$\geq \dfrac{\frac{n}{2} - (\frac{n}{2}+1)(1-\alpha)}{\frac{n}{2}+1}$

$= \dfrac{n - (n+2) + \alpha(n+2)}{n+2}$

$= \alpha - \frac{2}{n+2}$

$P(Y_{n+1} \in C(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2}$ □

remark

- More efficient and lower memory consumption.
- approximate in-sample coverage guarantee.

Theorem (theorem 3)

Under the conditions of Theorem 2, there is an absolute constant $c > 0$ such that, for any $\varepsilon > 0$,

$$P\left(\left|\frac{2}{n}\sum_{i\in I_2}\mathbf{1}\{Y_i \in C_{split}(X_i)\} - (1-\alpha)\right| \geq \varepsilon\right) \leq 2\exp\left(-cn^2\left(\varepsilon - \frac{4}{n}\right)_+^2\right)$$

- Split conformal inference can also be implemented using an unbalanced split.
  $|I_1| = \rho n$ and $|I_2| = (1-\rho)n$ for some $\rho \in (0,1)$

Multiple Splits

- Spliting introduces extra randomness into the procedure. One way to reduce this extra randomness is to combine inferences from several splits.
- split conformal prediction intervals $C_{\text{split},1}, \ldots, C_{\text{split},N}$ where each interval is constructed at level $1 - \alpha/N$.

$$C^{(N)}_{\text{split}}(x) = \bigcap_{j=1}^{N} C_{\text{split},j}(x),$$

over $x \in \mathbb{R}^d$.

- the prediction band $C^{(N)}_{\text{split}}(x)$ has marginal coverage level at least $1 - \alpha$

Remark

- decreases the variability from splitting
- cost :it is possible that the length of $C^{(N)}_{\text{split}}(x)$ grows with N, but taking an intersection reduces the size of the final interval(Bonferroni-intersection tradeoff.)
- Theorem 2.4. shows as $n \to \infty$, $C^{\text{N}}_{\text{split}}(X)$ is wider than $C_{\text{split}}(X)$.

Jackknife Prediction Intervals

- This method uses the quantiles of leave-one-out residuals to define prediction intervals
- The computational complexities is between full and split conformal methods
- Not guaranteed to have valid coverage in finite samples
- Finite-sample in-sample coverage property by symmetry :
  $P(Y_i \in C_{\text{jack}}(X_i)) \geq 1 - \alpha, \quad \text{for all } i = 1, \ldots, n.$
- Fragile for out-of-sample coverage (true predictive inference)

Algorithm 3

Algorithm 3 Jackknife Prediction Band

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$

Output: Prediction band, over $x \in \mathbb{R}^d$

for $i \in \{1, \ldots, n\}$ **do**

$\widehat{\mu}^{(-i)} = \mathcal{A}\big(\{(X_\ell, Y_\ell) : \ell \neq i\}\big)$

$R_i = |Y_i - \widehat{\mu}^{(-i)}(X_i)|$

end for

$d =$ the kth smallest value in $\{R_i : i \in \{1, \ldots, n\}\}$, where $k = \lceil n(1-\alpha) \rceil$

Return $C_{\text{jack}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

In-Sample Split Conformal Inference

- Goal: in-sample predictive inference, i.e evaluate prediction bands at some or all of the observed points $X_i, i = 1, \cdots, n$,
- One way is to treat each X_i as a new feature value and use the other n -1 points as the original features.
- Drawbacks
  - Degrades the computational efficiency
  - Not easy to show empirical coverage $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \in C(X_i)\}$ to be at least $1 - \alpha$
- **rank-one-out or ROO split conformal inference**. Similar to split conformal, but the ranking is conducted in a leave-one-out manner.

Algorithm 4

Algorithm 4 Rank-One-Out Split Conformal

Input: Data (X_i, Y_i), $i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$
Output: Prediction intervals at each X_i, $i = 1, \ldots, n$
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1$, $\mathcal{I}_2$
for $k \in \{1, 2\}$ **do**
  $\widehat{\mu}_k = \mathcal{A}\big(\{(X_i, Y_i) : i \in \mathcal{I}_k\}\big)$
  for $i \notin I_k$ **do**
    $R_i = |Y_i - \widehat{\mu}_k(X_i)|$
  end for
  for $i \notin I_k$ **do**
    $d_i =$ the mth smallest value in $\{R_j : j \notin \mathcal{I}_k,\ j \neq i\}$, where $m = \lceil n/2(1-\alpha) \rceil$
    $C_{\text{roo}}(X_i) = [\widehat{\mu}_k(X_i) - d_i, \widehat{\mu}_k(X_i) + d_i]$
  end for
end for
Return intervals $C_{\text{roo}}(X_i)$, $i = 1, \ldots, n$

ROO split conformal inference.

- Finite-sample in-sample coverage property:
  $P(Y_i \in C_{\text{roo}}(X_i)) \geq 1 - \alpha, \quad \text{for all } i = 1, \ldots, n.$
- Empirical in-sample average coverage $\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \in C(X_i)\}$
  Theorem 5.1
- A simpler, and conservative approximation to each in-sample prediction interval $C_{\text{roo}}(X_i))$ is
  $\widetilde{C}_{\text{roo}}(X_i) = \left[\widetilde{\mu}_k(X_i) - \widetilde{d}_k, \widetilde{\mu}_k(X_i) + \widetilde{d}_k\right]$ where, using the notation of Algorithm 4. Define $\widetilde{d}_k$ to be the mth smallest element of the set $\{R_i : i \in I_k\}$, for $m = \lceil (1-\alpha)n/2 \rceil + 1$.

Locally-Weighted Conformal Inference

- The full conformal and split conformal methods both tend to produce prediction bands C(x) whose width is roughly constant over $X \in \mathbb{R}^d$
- If the residual variance varies with X, we want the conformal band that can account for nonconstant residual variance.
- Recall the conformal inference method will have valid coverage if the conformity score function is symmetric in its first n arguments.
- Scaling the fitted residuals inversely by an estimated error spread

$$R_{y,i} = \frac{|Y_i - \hat{\mu}_y(X_i)|}{\hat{\rho}_y(X_i)}, \quad i = 1, \ldots, n, \quad \text{and} \quad R_{y,n+1} = \frac{|y - \hat{\mu}_y(x)|}{\hat{\rho}_y(x)}$$

Locally-Weighted Conformal Inference

- Scaling the fitted residuals inversely by an estimated error spread

$$R_{y,i} = \frac{|Y_i - \hat{\mu}_y(X_i)|}{\hat{\rho}_y(X_i)}, \quad i = 1, \ldots, n, \quad \text{and} \quad R_{y,n+1} = \frac{|y - \hat{\mu}_y(x)|}{\hat{\rho}_y(x)}$$

  - $\hat{\rho}_y(x)$ denotes an estimate of the conditional mean absolute deviation (MAD).
  - Exits in more cases than standard deviation.
- For the split conformal and the ROO split conformal methods:

$$R_{y,i} = \frac{|Y_i - \hat{\mu}(X_i)|}{\hat{\rho}(X_i)}$$

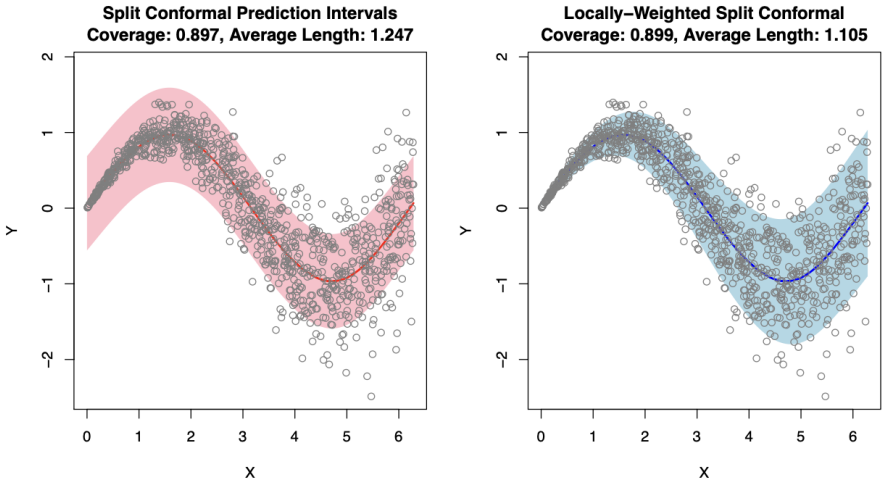


Figure: heteroskedasticity

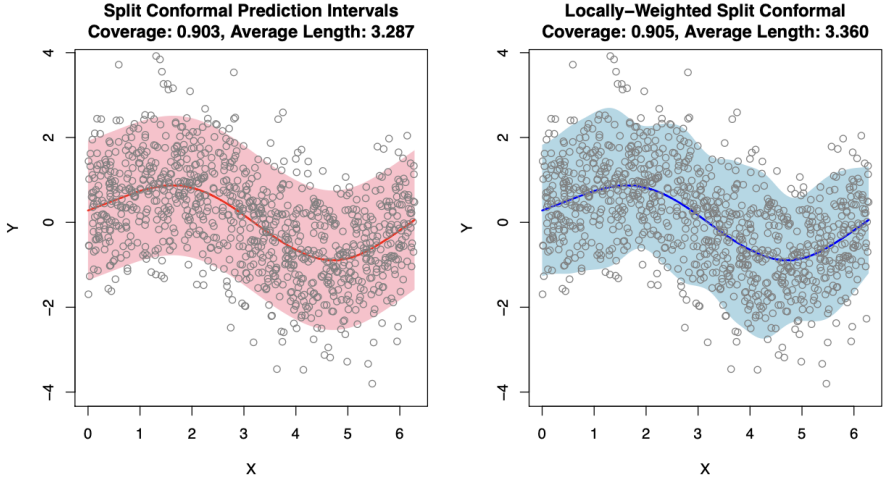


Figure: without heteroskedasticity

Figure: (For each group)The top left panel shows the split conformal band, and the top right shows the locally weighted split conformal band; The bottom left and right panels plot the empirical local coverage and local length measures

Statistical Accuracy

- Comparison to a "super oracle" and a regular oracle:
  - The super oracle has complete knowledge of the regression function and the error distribution
  - The regular oracle has knowledge only of the residual distribution.
- Three main theoretical results are:
  - If the base estimator is consistent, then the two oracle bands have similar lengths
  - If the base estimator is stable under resampling and small perturbations, then the conformal prediction bands are close to the oracle band.
  - If the base estimator is consistent, then the conformal prediction bands are close to the super oracle

Statistical Accuracy

Assumption A0 (iid data). We observe iid data (X_i, Y_i), $i = 1, \ldots, n$ from a common distribution P on $\mathbb{R}^d \times \mathbb{R}$, with mean function $\mu(x) = \mathbb{E}(Y \mid X = x), x \in \mathbb{R}^d$.

Assumption A1 (Independent and symmetric noise). For $(X, Y) \sim P$, the noise variable $\epsilon = Y - \mu(X)$ is independent of X, and the density function of ϵ is symmetric about 0 and nonincreasing on $[0, \infty)$.

Assumption A2 (Sampling stability). For large enough n,

$$\mathbb{P}(\|\widehat{\mu}_n - \widetilde{\mu}\|_\infty \geq \eta_n) \leq \rho_n,$$

for some sequences satisfying $\eta_n = o(1)$, $\rho_n = o(1)$ as $n \to \infty$, and some function $\widetilde{\mu}$.

Assumption A3 (Perturb-one sensitivity). For large enough n,

$$\mathbb{P}\left(\sup_{y \in \mathcal{Y}} \|\widehat{\mu}_n - \widehat{\mu}_{n,(X,y)}\|_\infty \geq \eta_n\right) \leq \rho_n,$$

for some sequences satisfying $\eta_n = o(1)$, $\rho_n = o(1)$ as $n \to \infty$.

Assumption A4 (Consistency of base estimator). For n large enough,

$$\mathbb{P}\left(\mathbb{E}_X[(\widehat{\mu}_n(X) - \mu(X))^2 \mid \widehat{\mu}_n] \geq \eta_n\right) \leq \rho_n,$$

for some sequences satisfying $\eta_n = o(1)$, $\rho_n = o(1)$ as $n \to \infty$.

Empirical Study: Comparisons to Parametric Intervals for Linear Regression

Table: Comparison of prediction intervals in low-dimensional problems with $n = 100$, $d = 10$.

Setting A

	Conformal	Jackknife	Split	Parametric
Coverage	0.904 (0.005)	0.892 (0.005)	0.905 (0.008)	0.9 (0.006)
Length	3.529 (0.044)	3.399 (0.04)	3.836 (0.082)	3.477 (0.036)
Time	1.106 (0.004)	0.001 (0)	0 (0)	0.001 (0)

Setting B

	Conformal	Jackknife	Split	Parametric
Coverage	0.915 (0.005)	0.901 (0.006)	0.898 (0.006)	0.933 (0.007)
Length	6.594 (0.254)	6.266 (0.254)	7.384 (0.532)	8.714 (0.768)
Time	1.097 (0.002)	0.001 (0)	0.001 (0)	0.001 (0)

Setting C

	Conformal	Jackknife	Split	Parametric
Coverage	0.904 (0.004)	0.892 (0.005)	0.896 (0.008)	0.943 (0.005)
Length	20.606 (1.161)	19.231 (1.082)	24.882 (2.224)	33.9 (4.326)
Time	1.105 (0.002)	0.001 (0)	0.001 (0)	0 (0)

Empirical Study: Comparisons to Parametric Intervals for Linear Regression

Table: Comparison of prediction intervals in high-dimensional problems with $n = 500$, $d = 490$.

Setting A

	Conformal	Jackknife	Parametric
Coverage	0.918 (0.01)	0.894 (0.014)	0.913 (0.013)
Length	7.431 (0.164)	27.863 (0.835)	28.423 (0.886)
Time	99.541 (0.2)	0.926 (0.003)	0.391 (0.001)

Setting B

	Conformal	Jackknife	Parametric
Coverage	0.88 (0.014)	0.874 (0.016)	0.878 (0.017)
Length	56.74 (10.097)	71.084 (8.769)	72.727 (9.087)
Time	96.31 (0.207)	0.912 (0.002)	0.386 (0.001)

Setting C

	Conformal	Jackknife	Parametric
Coverage	0.894 (0.014)	0.887 (0.015)	0.898 (0.015)
Length	233.361 (23.142)	266.458 (21.238)	275.388 (24.49)
Time	171.041 (0.443)	1.527 (0.008)	0.514 (0.003)

Empirical Study: Comparisons to Parametric Intervals for Linear Regression

Table: Comparison of prediction intervals in high-dimensional problems with $n = 500$, $d = 490$, using ridge regularization.

Setting A

	Conformal	Jackknife	Split	Parametric
Coverage	0.908 (0.004)	0.907 (0.004)	0.911 (0.004)	1 (0)
Length	3.354 (0.017)	3.329 (0.017)	3.379 (0.027)	26.697 (0.117)
Test error	0.996 (0.019)	0.996 (0.019)	0.995 (0.019)	0.996 (0.019)
Time	99.541 (0.2)	0.926 (0.003)	0.184 (0.001)	0.391 (0.001)

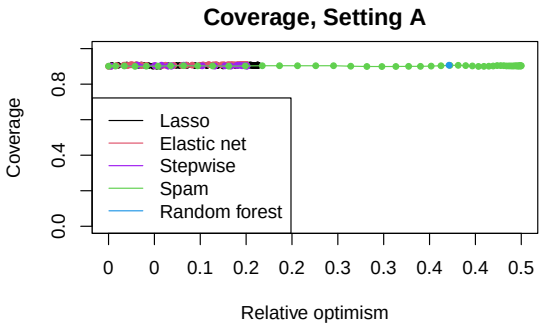
Setting B

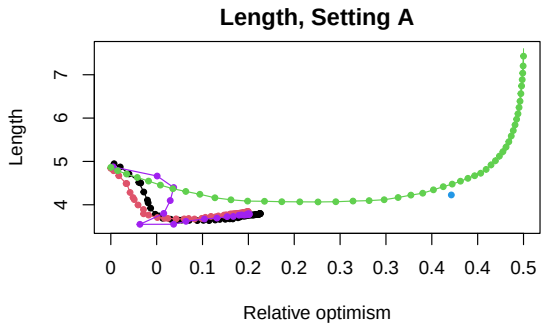
	Conformal	Jackknife	Split	Parametric
Coverage	0.909 (0.005)	0.904 (0.005)	0.904 (0.005)	1 (0)
Length	6.058 (0.163)	5.762 (0.075)	5.807 (0.079)	74.652 (8.306)
Test error	5.789 (0.755)	5.789 (0.755)	5.786 (0.768)	5.789 (0.755)
Time	96.31 (0.207)	0.912 (0.002)	0.182 (0)	0.386 (0.001)

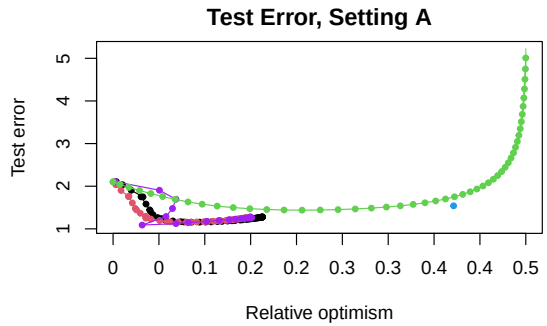
Setting C

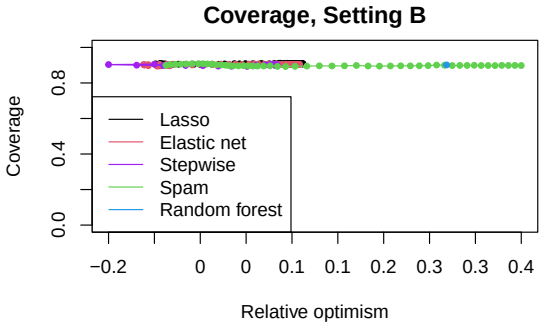
	Conformal	Jackknife	Split	Parametric
Coverage	0.907 (0.005)	0.899 (0.005)	0.903 (0.004)	0.999 (0)
Length	15.791 (0.328)	14.586 (0.224)	15.206 (0.313)	266.225 (16.154)
Test error	104.053 (24.982)	104.053 (24.982)	104.192 (25.087)	104.053 (24.982)
Time	171.041 (0.443)	1.527 (0.008)	0.2 (0.002)	0.514 (0.003)

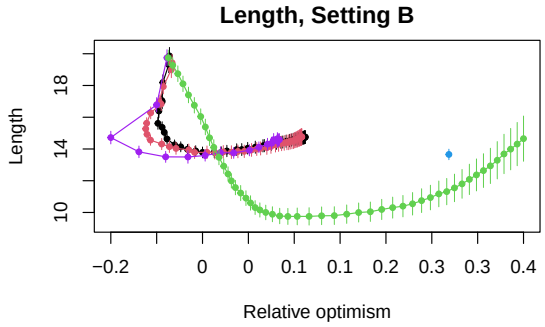
Empirical Study: Comparisons of Conformal Intervals Across Base Estimators(low dim)

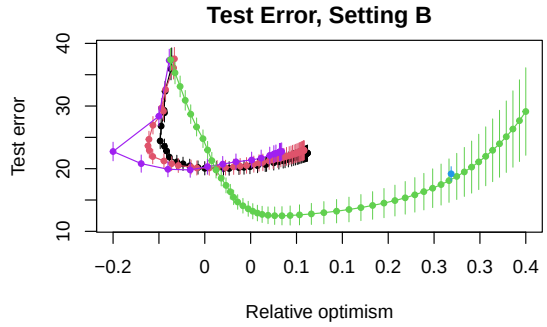


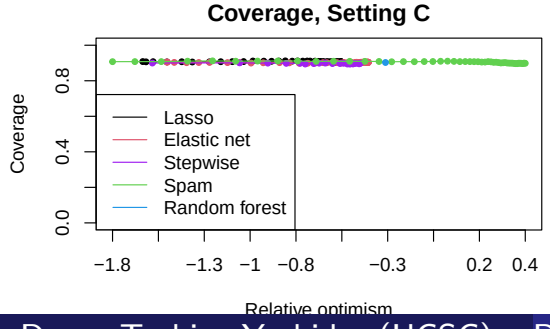


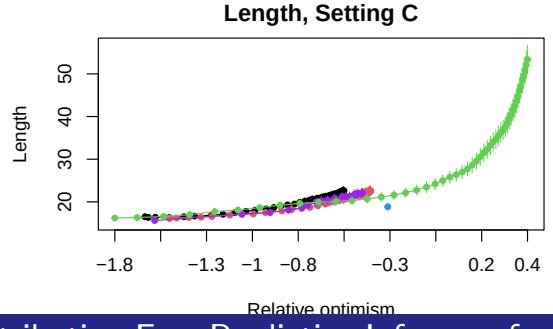


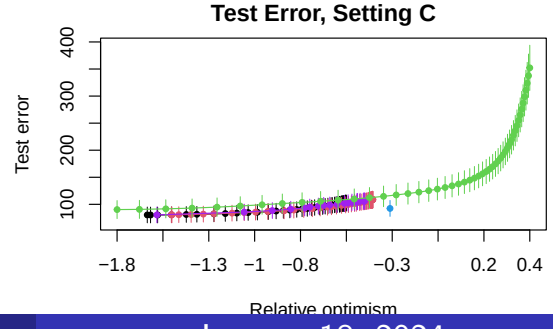




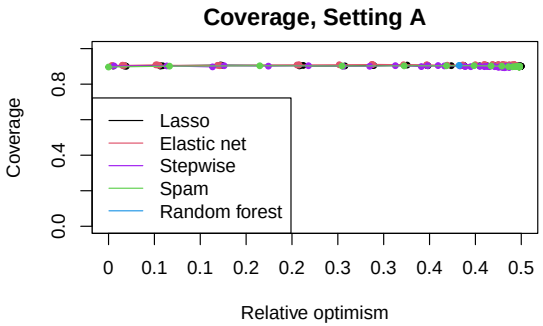


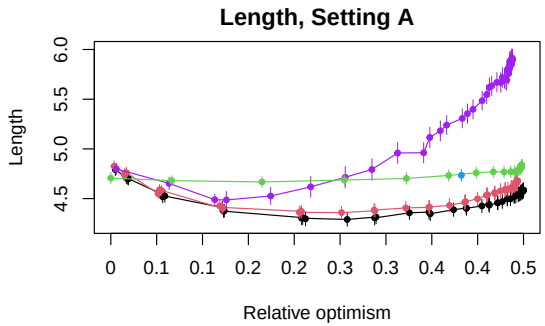


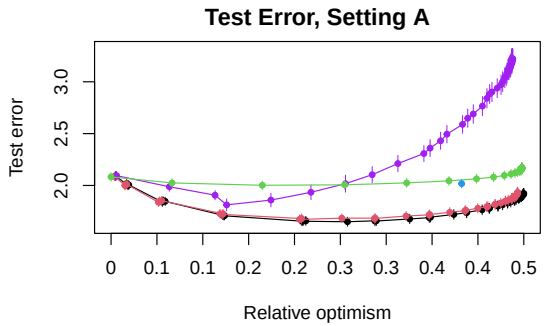


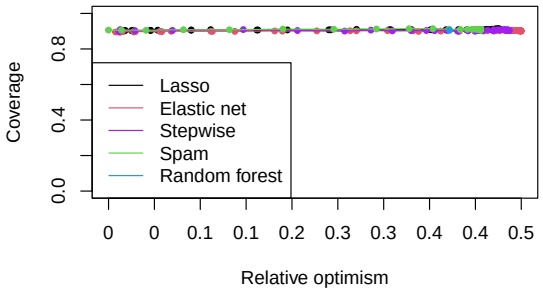


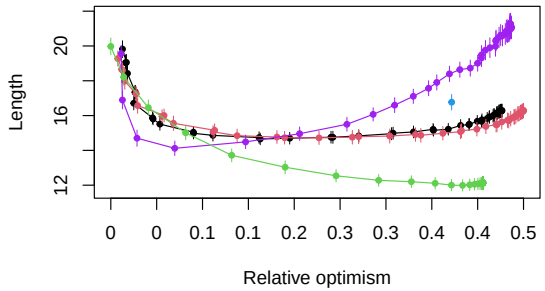
Empirical Study: Comparisons of Conformal Intervals Across Base Estimators(high dim)

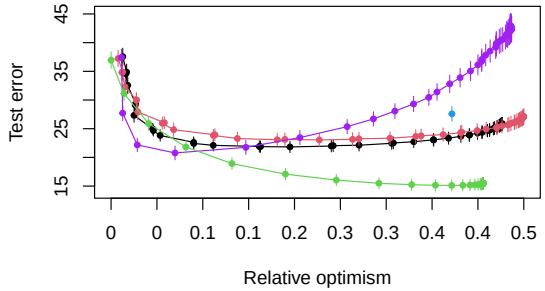


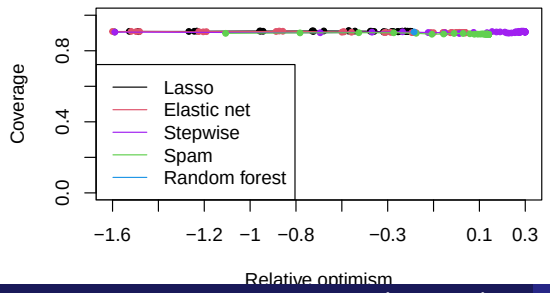


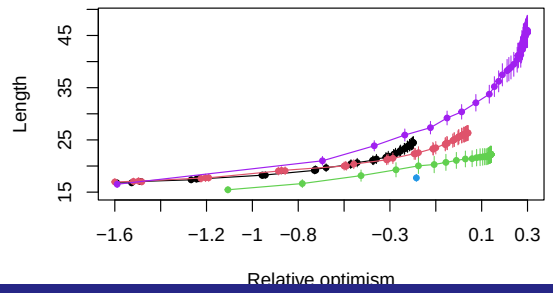


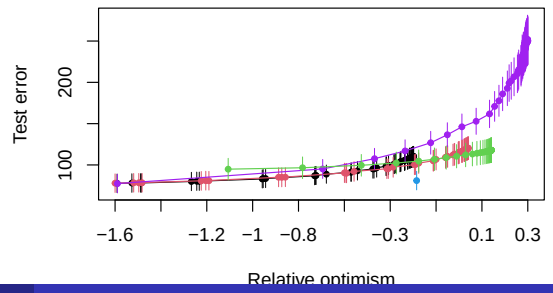












Model-Free prediction-based approaches for inferring variable importance

- leave-one-covariate-out
- $\hat{\mu}$: fit on data $(X_i, Y_i), i \in I_1$ for some $I_1 \subseteq \{1, \ldots, n\}$.
- $\hat{\mu}_{(-j)}$: refit on the data set $(X_i(-j), Y_i), i \in I_1$, where in each $X_i(-j) = (X_i(1), \ldots, X_i(j-1), X_i(j+1), \ldots X_i(d)) \in \mathbb{R}^{d-1}$.
- $\Delta_j(X_{n+1}, Y_{n+1}) = |Y_{n+1} - \hat{\mu}_{(-j)}(X_{n+1})| - |Y_{n+1} - \hat{\mu}(X_{n+1})|$
  - Excess prediction error of covariate j, at a new i.i.d. draw (X_{n+1}, Y_{n+1})
  - a random variable that measures the increase in prediction error due to not having access to covariate j in data set.

Local Measure of Variable Importance

- Let C denote a conformal prediction with coverage $1-\alpha$
- Define prediction set
  $W_j(x) = \left\{|y - \hat{\mu}_{(-j)}(x)| - |y - \hat{\mu}(x)| : y \in C(x)\right\}$
- From the finite-sample validity of C,

$$P\left(\Delta_j(X_{n+1}, Y_{n+1}) \in W_j(X_{n+1}) \text{ for all } j = 1, \ldots, d\right) \geq 1 - \alpha. \quad (15)$$

- Marginal coverage over X_{n+1}, does not hold conditionally at $X_{n+1} = x$, loosely use intervals $W_j(x)$ to show effect of covariate j locally.
- Example: d=6. with mean function
  $\mu(x) = \sum_{j=1}^{6} f_j(x(j))$, with $f_1(t) = \sin(\pi(1+t))\mathbf{1}\{t < 0\}, f_2(t) = \sin(\pi t), f_3(t) = \sin(\pi(1+t))\mathbf{1}\{t > 0\}$, and $f_4 = f_5 = f_6 = 0$.
- $n = 1000$ i.i.d pairs $(X_i, Y_i), i = 1, \ldots, 1000$, where each $X_i \sim \text{Unif}[-1, 1]^d$ and $Y_i = \mu(X_i) + \varepsilon_i$ for $\varepsilon_i \sim \mathcal{N}(0, 1)$.

Local Measure of Variable Importance

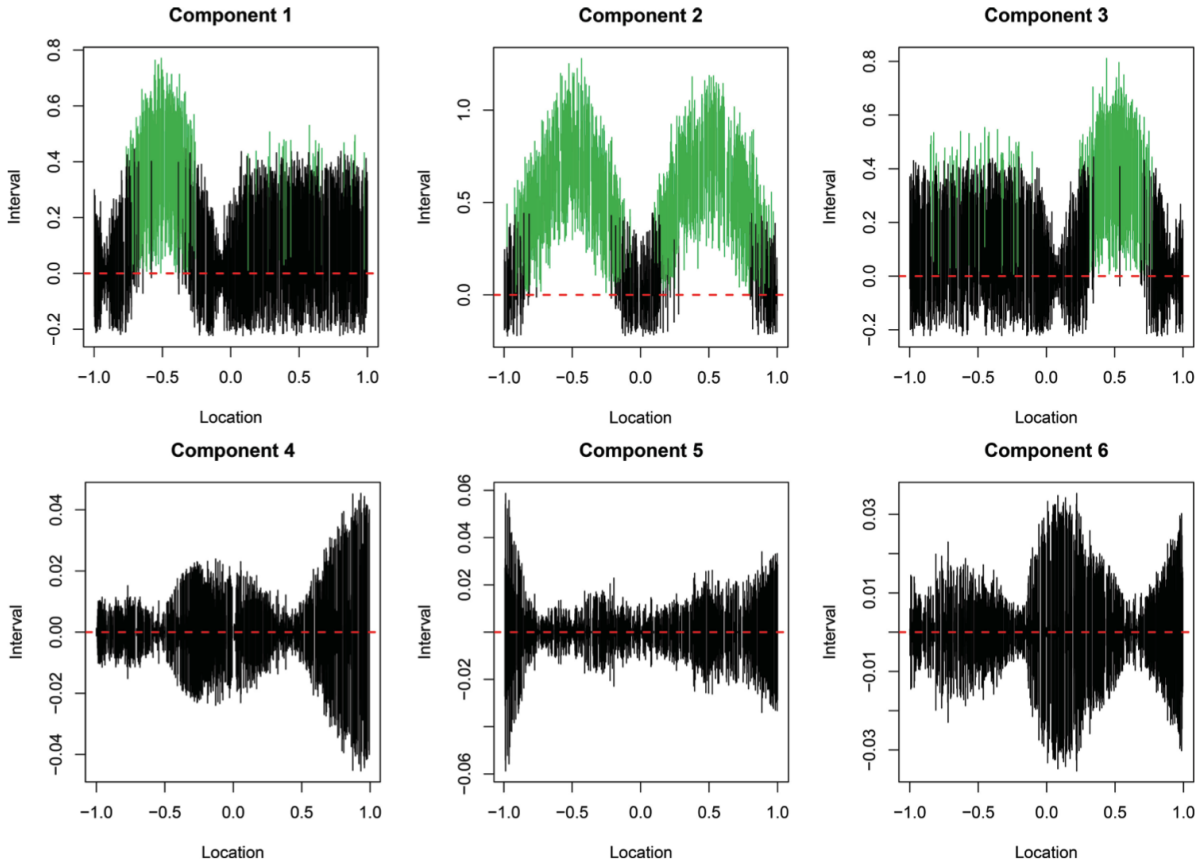


Figure: interval $W_j(X_{n+1})$ using the ROO split conformal technique at the miscoverage level $\alpha = 0.1$,

Global Measures of Variable Importance

- Consider splitting approach to train $\hat{\mu}$ and $\hat{\mu}_{(-j)}$ is $I_1 \subseteq \{1, \ldots, n\}$
- Denote by I_2 its complement, and by $D_k = \{(X_i, Y_i) : i \in I_k\}, k = 1, 2$ the data samples in each index set
- Define $G_j(t) = \mathbb{P}(\Delta_j(X_{n+1}, Y_{n+1}) \leq t | D_1), t \in \mathbb{R}$,
- Infer parameters of G_j such as its mean or median.

Global Measures of Variable Importance

- Mean: $\theta_j = \mathbb{E}\left[\Delta_j(X_{n+1}, Y_{n+1})|D_1\right]$
  - An asymptotic $1-\alpha$ confidence interval

$$\left[\hat{\theta}_j - \frac{z_{\alpha/2}\hat{\sigma}_j}{\sqrt{n/2}}, \hat{\theta}_j + \frac{z_{\alpha/2}\hat{\sigma}_j}{\sqrt{n/2}}\right]$$

  - $\hat{\theta}_j = (n/2)^{-1}\sum_{i\in I_2}\Delta_j(X_i, Y_i)$. S^2 sample variance measured on D_2
- Median: $m_j = median\left[\Delta_j(X_{n+1}, Y_{n+1})|D_1\right]$,
  - Nonparametric tests such as the sign test or the Wilcoxon signed-rank test, applied to $\Delta_j(X_{n+1}, Y_{n+1})$, i $\in I_2$.
  - $H_0 : m_j \leq 0 \quad$ versus $\quad H_1 : m_j > 0$

References

- Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman (2018). Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association, 113(523), 1094-1111.