Distribution-Free Predictive Inference For Regression

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman

Kiho Park
parkkiho@uchicago.edu

1 Introduction

Consider i.i.d. regression data $Z_1, \cdots, Z_n \sim P$ where $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ and $X_i = (X_i(1), \cdots, X_i(d))$. Let $\mu(x) = \mathbb{E}(Y \mid X = x)$ denote the regression function. Given a nominal miscoverage level $\alpha \in (0, 1)$, we want to construct a prediction band $C \subseteq \mathbb{R}^d \times \mathbb{R}$ satisfying

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

where $C(x) = \{y \in R : (x, y) \in C\}$. The our goal is to construct the bands that guarantee finite-sample validity and state the importance of each covariate in the prediction model for Y_{n+1} given X_{n+1}.

The conformal prediction framework is invented by Vovk et al. (2005). We consider testing the null hypothesis that $Y_{n+1} = y$ and construct a valid p-value based on the empirical quantiles of the augmented sample which makes the procedure immune to overfitting. Then, we decide if a value y is to be included in $C(X_{n+1})$. Conformal inference has been studied as a batch method, in various settings. However, there is very little work on prediction sets in high-dimensional regression. We provides conformal inference in the context of high-dimensional regression, new theoretical insights, approximation of certain oracle methods, and empirical study. Also, we develop two extensions of conformal inference: prediction intervals with in-sample coverage, and prediction intervals with varying local length. Lastly, we propose new, model-free, prediction-based approaches for inferring variable importance.

2 Conformal Inference

Let $U_1, \cdots, U_n$ be i.i.d. samples of a random variable. We define the sample quantile $\widehat{q}_{1-\alpha}$ based on $U_1, \cdots, U_n$ by

$$\widehat{q}_{1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)} & \text{if } \lceil (n+1)(1-\alpha) \rceil \leq n \\ \infty & \text{otherwise} \end{cases}$$

where $U_{(1)} \leq \cdots \leq U_{(n)}$ is the order statistics. By exchangeability,

$$\mathbb{P}(U_{n+1} \leq \widehat{q}_{1-\alpha}) \geq 1 - \alpha$$

In our regression problem, the prediction interval is defined by

$$C_{\text{naive}}(X_{n+1}) = \left[\widehat{\mu}(X_{n+1}) - \widehat{F}_n^{-1}(1-\alpha),\, \widehat{\mu}(X_{n+1}) + \widehat{F}_n^{-1}(1-\alpha)\right]$$

where $\widehat{\mu}$ is an estimator of the underlying regression function and $\widehat{F}_n^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of $\widehat{F}_n$ which is the empirical distribution of the fitted residuals $|Y_i - \widehat{\mu}(X_i)|$. This naive method can grossly undercover since the fitted residual distribution can often be biased. Conformal prediction intervals overcome the deficiencies of the naive intervals.

2.1 Conformal Prediction Sets

The conformal prediction interval is computed by the following algorithm.

Algorithm 1 Conformal Prediction

```
Input: Data $(X_i, Y_i)$, $i = 1, \cdots, n$, miscoverage level $\alpha \in (0,1)$, regression algorithm $\mathcal{A}$, points $\mathcal{X}_{\text{new}} = \{X_{n+1}, X_{n+2}, \cdots\}$ at which to construct prediction intervals, and values $\mathcal{Y}_{\text{trial}} = \{y_1, y_2, \cdots\}$ 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
        $\widehat{\mu}_y = \mathcal{A}(\{(X_1, Y_1), \cdots, (X_n, Y_n), (x, y)\})$
        $R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|$, $i = 1, \cdots, n$ 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 $x \in \mathcal{X}_{\text{new}}$
```

The conformal prediction interval $C_{\text{conf}}(X_{n+1})$ has valid finite-sample coverage and it is also accurate as described by the following theorem.

Theorem 2.1. *If (X_i, Y_i), $i = 1, \cdots, n$ are i.i.d., then for a new i.i.d. pair (X_{n+1}, Y_{n+1}),*

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

for the conformal prediction band $C_{conf}(X_{n+1})$. In addition, if $R_{y,i}$ have a continuous joint distribution,

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

Note that the conformal prediction interval guarantees average (or marginal) coverage, not conditional coverage, i.e., $\mathbb{P}(Y_{n+1} \in C(x) \mid X_{n+1} = x) \geq 1 - \alpha$ for all $x \in \mathbb{R}^d$.

2.2 Split Conformal Prediction Sets

The original conformal prediction method is computationally intensive. There is an alternative approach: split conformal prediction (or inductive conformal inference).

Algorithm 2 Split Conformal Prediction

```
Input: Data $(X_i, Y_i)$, $i = 1, \cdots, n$, miscoverage level $\alpha \in (0,1)$, regression algorithm $\mathcal{A}$
Output: Predictions band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \cdots, n\}$ 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 $k$th 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), $i = 1, \cdots, n$ are i.i.d., then for a new i.i.d. pair (X_{n+1}, Y_{n+1}),*

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

for the conformal prediction band $C_{conf}(X_{n+1})$. In addition, if R_i, $i \in \mathcal{I}_2$ have a continuous joint distribution,

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

This split conformal inference is extremely efficient and can hold an advantage in terms of memory requirements. The following theorem describes that the split conformal prediction intervals provides an approximate in-sample coverage guarantee.

Theorem 2.3. *Under the conditions of Theorem 2.2, there is an absolute constant $c > 0$ such that, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{2}{n}\sum_{i \in \mathcal{I}_2} \mathbb{1}\{Y_i \in C_{split}(X_i)\} - (1-\alpha)\right| \geq \epsilon\right) \leq 2\exp\left(-cn^2(\epsilon - 4/n)_+^2\right)$$

2.3 Multiple Splits

Splitting improves the speed, but introduces extra randomness into the procedure. To solve this problem, we can split the training data N times and construct $C_{\text{split},1}, \cdots, C_{\text{split},N}$ at level $1 - \alpha/N$. Then, we define

$$C_{\text{split}}^{(N)}(x) = \bigcap_{j=1}^{N} C_{\text{split},j}(x), \quad \text{over } x \in \mathbb{R}^d$$

There is a “Bonferroni-intersection tradeoff” in this method, and we get larger intervals as N increases. Therefore, we suggest using a single split.

2.4 Jackknife Prediction Intervals

Jackknife has the computational complexity between that of the full and split conformal methods.

Algorithm 3 Jackknife Prediction Band

```
Input: Data $(X_i, Y_i)$, $i = 1, \cdots, n$, miscoverage level $\alpha \in (0,1)$, regression algorithm $\mathcal{A}$
Output: Predictions band, over $x \in \mathbb{R}^d$
for $i \in \{1, \cdots, n\}$ do
    $\widehat{\mu}^{(-i)} = \mathcal{A}(\{(X_\ell, Y_\ell) : \ell \neq i\})$
    $R_i = |Y_i - \widehat{\mu}^{(-i)}(X_i)|$
end for
$d =$ the $k$th smallest value in $\{R_i, i \in \{1, \cdots, n\}\}$ where $k = \lceil n(1-\alpha) \rceil$
Return $C_{\text{jack}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$
```

The jackknife method utilizes more of the training data when constructing the absolute residuals than the split conformal method. It would produce shorter intervals. However, its prediction intervals are not guaranteed to have valid coverage in finite samples. Some conditions are needed for the finite sample validity, even asymptotically.

3 Statistical Accuracy

We investigate the statistical accuracy of conformal prediction intervals by bounding the length of the resulting intervals $C(X)$. We need the following assumptions.

Assumption 1 (i.i.d. data)**.** *We observe i.i.d. data* (X_i, Y_i)*,* $i = 1, \cdots, n$ *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 2 (Independent and symmetric noises)**.** *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* $[0, \infty)$*.*

We compare the lengths of prediction bands constructed with the full and split conformal methods to the length of the idealized prediction bands obtained by two oracles: the "super oracle" and a regular oracle.

Under the assumptions, the super oracle prediction band is

$$C_s^*(x) = [\mu(x) - q_\alpha, \mu(x) + q_\alpha]$$

where q_α is the α upper quantile of $\mathcal{L}(|\epsilon|)$. This band is the shortest length among all bands with conditional coverage ($\mathbb{P}(Y \in C_s^*(x) \mid X = x) \geq 1 - \alpha$) and marginal coverage.

The regular oracle is

$$C_o^*(x) = [\widehat{\mu}_n(x) - q_{n,\alpha}, \widehat{\mu}_n(x) + q_{n,\alpha}]$$

where $q_{n,\alpha}$ is the α upper quantile of $\mathcal{L}(|Y - \widehat{\mu}_n(X)|)$. This band only offers marginal coverage ($\mathbb{P}(Y \in C_o^*(x)) \geq 1 - \alpha$) over the randomness of the $(n + 1)$ pairs.

3.1 Comparing the Oracles

Let $\Delta_n(x) = \widehat{\mu}_n(x) - \mu(x)$ denote the estimation error. The the following theorem holds.

Theorem 3.1 (Comparing the oracle bands)**.** *Under Assumptions 1 and 2, let* F, f *be the distribution and density functions of* $|\epsilon|$*. Assume further that* f *has continuous derivative that is uniformly bounded by* $M > 0$*. Let* F_n, f_n *be the distribution and density functions of* $|Y - \widehat{\mu}_n(X)|$*. Then we have*

$$\sup_{t>0} |F_n(t) - F(t)| \leq \frac{M}{2} \mathbb{E}\Delta_n^2(X)$$

where the expectation is taken over the randomness of $\widehat{\mu}_n$ *and* X*. Moreover, if* f *is lower bounded by* $r > 0$ *on* $(q_\alpha - \eta, q_\alpha + \eta)$ *for some* $\eta > \frac{M}{2r} \mathbb{E}\Delta_n^2(X)$*, then*

$$|q_{n,\alpha} - q_\alpha| \leq \frac{M}{2r} \mathbb{E}\Delta_n^2(X)$$

This theorem means that the two oracle bands have similar lengths if the base estimator is consistent (even just small $\mathbb{E}\Delta_n^2(X)$ is enough).

3.2 Oracle Approximation Under Stability Assumptions

If the base estimator is stable under resampling and small perturbations, then the split and full conformal prediction bands are close to the regular oracle band. The following theorems describe it.

3.2.1 Case I: Split conformal

Assumption 3 (Sampling stability). *For large enough n,*

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

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

Theorem 3.2 (Split conformal approximation of oracle). *Fix $\alpha \in (0,1)$, and let $C_{n,split}$ and $\nu_{n,split}$ denote the split conformal interval and its width. Under Assumptions 1, 2, 3, assume further that $\tilde{f}$, the density of $|Y - \tilde{\mu}(X)|$, is lower bounded away from zero in an open neighborhood of its α upper quantile. Then*

$$\nu_{n,split} - 2q_{n,\alpha} = O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2})$$

3.2.2 Case II: Full conformal

Assumption 4 (Perturb-one sensitivity). *Let $\widehat{\mu}_{n,(X,y)}$ denote the fitted regression function with augmented data with $(n+1)$st data point (X,y). 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 $n \to \infty$.

Theorem 3.3 (Full conformal approximation of oracle). *Under the same assumptions as in Theorem 3.2, assume in addition that Y is supported on $\mathcal{Y}$ such that Assumption 4 holds. Fix $\alpha \in (0,1)$, and let $C_{n,conf}(X)$ and $\nu_{n,conf}(X)$ denote the conformal interval and its width at X. Then*

$$\nu_{n,conf}(X) - 2q_{n,\alpha} = O_{\mathbb{P}}(\rho_n + \eta_n + n^{-1/2})$$

3.3 Super Oracle Approximation Under Consistency Assumptions

Definition 1 (Asymptotic conditional coverage). *We say that a sequence C_n of (possibly) random prediction bands has asymptotic conditional coverage at the level $(1-\alpha)$ if there exists a sequence of (possibly) random sets $\Lambda_n \subseteq \mathbb{R}^d$ such that $\mathbb{P}(X \in \Lambda_n \mid \Lambda_n) = 1 - o_{\mathbb{P}}(1)$ and*

$$\inf_{x \in \Lambda_n} |\mathbb{P}(Y \in C_n(x) \mid X = x) - (1-\alpha)| = o_{\mathbb{P}}(1)$$

Assumption 5 (Consistency of base estimator). *For n large enough,*

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

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

Theorem 3.4 (Split conformal approximation of super oracle). *Under Assumptions 1, 2, 5, assuming in addition that $|Y - \mu(X)|$ has density bounded away from zero in an open neighborhood of its α upper quantile, the split conformal interval satisfies*

$$L(C_{n,split}(X) \triangle C_s^*(X)) = o_{\mathbb{P}}(1)$$

where $L(A)$ denotes the Lebesgue measure of a set A, and $A \triangle B$ denotes the symmetric difference between sets A, B. Thus, $C_{n,split}$ has asymptotic conditional coverage at the level $1-\alpha$.

Theorem 3.5 (Full conformal approximation of super oracle)**.** *Under the same conditions as in Theorem 3.4, and in addition Assumption 4, we have*

$$L(C_{n,conf}(X) \triangle C_s^*(X)) = o_{\mathbb{P}}(1)$$

and thus $C_{n,conf}$ has asymptotic conditional coverage at the level $1-\alpha$.

These theorems mean that the conformal prediction bands are close to the super oracle if the base estimator is consistent.

3.4 A High-dimensional Sparse Regression Example

Consider the data are i.i.d. replicates from the model $Y = X^T\beta + \epsilon$, with ϵ being independent of X with mean 0 and variance σ^2. Assume $X \in [-1,1]^d$ and $\epsilon \in [-R,R]$ and $\beta \in \mathbb{R}^d$ has the support J with cardinality $s < \min\{n,d\}$ nonzero entries. We use the lasso

$$\widehat{\beta}_{n,\text{lasso}} = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda||\beta||_1$$

as the base estimator where $\lambda \geq 0$ is a tuning parameter.

For the split conformal method, assume that the covariance matrix Σ of X satisfies the restricted eigenvalue condition for $\kappa > 0$:

$$\min_{v:||v_J||_2=1,||v_{J^c}||_1\leq 3||v_J||_1} v^T\Sigma v \geq \kappa^2$$

Then, by Theorem 7.2 of Bickel et al. (2009), if $\lambda = C\sigma\sqrt{\log d/n}$, with probability at least $1-d^{-c}$,

$$||\widehat{\beta}_{n,\text{lasso}} - \beta||_1 \leq C'\kappa^2 Rs\sqrt{\log d/n}$$

Therefore, Assumptions 3 and 5 hold with $\tilde{\mu}(x) = x^T\beta$, $\eta_n = C'\kappa^2 Rs\sqrt{\log d/n}$, and $\rho_n = d^{-c}$.

For the full conformal method, assume there are a constant $\Phi \in (0,1/2)$ such that the absolute values of all nonzero entries of β are in $[\Phi, 1-\Phi]$, and a constant $\delta \in (0,1/4)$ such that $||\Sigma_{J^cJ}\Sigma_{JJ}^{-1}\text{sign}(\beta_J)||_\infty \leq 1/4-\delta$. Also, assume that the active block of the covariance matrix Σ_{JJ} has minimum eigenvalue $\Psi > 0$. Then, by Theorem 8 of Thakurta & Smith (2013),

$$||\widehat{\beta}_{J,\text{lasso}}(X,y) - \widehat{\beta}_{J,\text{ols}}(X,y)||_\infty \leq O_{\mathbb{P}}(n^{1/2}) + O_{\mathbb{P}}(\lambda) = O_{\mathbb{P}}(\sqrt{\log d/n})$$

Therefore, Assumptions 4 and 5 hold for any η_n such that $\eta_n\sqrt{n/\log d} \to \infty$, and $\rho_n = o(1)$.

4 Empirical Study

We consider three experimental settings: (A) linear, classical, (B) nonlinear, heavy-tailed, (C) linear, heteroskedastic, heavy-tailed, correlated features.

4.1 Comparisons to Parametric Intervals from Linear Regression

We compare the intervals by four methods: full conformal, Jackknife, split conformal, and classical parametric.

In the low-dimensional problem, all four methods give reasonable coverage for Setting A. For Setting B and C, the parametric intervals are wider. Therefore, the coverage provided by the conformal approach is robust to the model misspecification.

In the high-dimensional problems, the full conformal interval outperforms the parametric interval and the jackknife interval, even in Setting A. When we use ridge regression for the high-dimensional problems, the conformal methods outperform the parametric one. Moreover, the split conformal method exhibits a clear computational advantage compared to the full conformal method, with similar performance. Therefore, we can easily combine split conformal with computationally heavy estimators that involve cross-validation or bootstrap.

4.2 Comparisons to Conformal Intervals Across Base Estimators

We compare the conformal intervals across a variety of base estimators: forward stepwise regression, the lasso, the elastic net, sparse additive models (SPAM), and random forests.

In the low-dimensional problems, the best test errors are obtained by the linear methods (lasso, elastic net, stepwise) in Setting A, and by SPAM in Setting B. In Setting C, all estimators perform quite poorly. In the high-dimensional problems, the results are similar. Note that across all settings and estimators, the coverage of the conformal prediction intervals is almost exactly 90%, and the lengths are highly correlated with test errors.

5 Extensions of Conformal Inference

We describe some extensions of this framework to improve the interpretability and applicability of conformal inference.

5.1 In-Sample Split Conformal Inference

Given samples and a method, we would often like to evaluate the prediction band at some or all of the observed points X_i. However, if we apply the full or split conformal prediction methods at an observed feature value, then it is not easy to establish finite-sample validity of these methods.

We propose rank-one-out or ROO split conformal inference presented in Algorithm 4. By symmetry in the construction, the ROO split conformal intervals have the in-sample finite-sample coverage property

$$\mathbb{P}(Y_i \in C_{\mathrm{roo}}(X_i)) \geq 1-\alpha, \quad \text{for all } i = 1, \cdots, n$$

Also, the following theorem for the empirical in-sample average coverage holds by McDiarmid's inequality.

Theorem 5.1. *If (X_i, Y_i), $i = 1, \cdots, n$ are i.i.d., then for the ROO split conformal band C_{roo}, there is an absolute constant $c > 0$, such that for all $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \in C_{roo}(X_i)\} \geq 1-\alpha-\epsilon\right) \geq 1-2\exp(-cn\epsilon^2)$$

Moreover, if we assume additionally that the residuals R_i, $i = 1, \cdots, n$, have a continuous joint distribution, then for all $\epsilon > 0$,

$$\mathbb{P}\left(1-\alpha-\epsilon \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \in C_{roo}(X_i)\} \leq 1-\alpha+\frac{2}{n}+\epsilon\right) \geq 1-2\exp(-cn\epsilon^2)$$

Algorithm 4 Rank-One-Out Split Conformal

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

5.2 Locally-Weighted Conformal Inference

The residual variance can vary with X, so we want the conformal band to adapt correspondingly. We can actually use any conformity score function. We modify the definition of residuals in the full conformal algorithm as

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

where $\widehat{\rho}_y(x)$ denotes an estimate of the conditional mean absolute deviation (MAD) of $(Y - \mu(X)) \mid X = x$, as a function of $x \in \mathbb{R}^d$. It can be estimated jointly or using the estimated conditional mean $\widehat{\mu}_y$. For the split conformal and the ROO split conformal methods, the locally-weighted residual is

$$R_i = \frac{|Y_i - \widehat{\mu}(X_i)|}{\widehat{\rho}(X_i)}, \; i \in \mathcal{I}_2$$

where $\widehat{\mu}$ and $\widehat{\rho}$ are fit on the samples in $\mathcal{I}_1$.

By a simple example, we show that the locally-weighted conformal band outperforms the split conformal band in terms of the local coverage. Note that the locally-weighted conformal band will have generally a inflated length due to the additional randomness in estimating the conditional MAD. However, in Appendix, we can check that there is not too much inflation in the length for homoskedasitc noise setting.

6 Model-Free Variable Importance: LOCO

How do we assess variable importance when we are treating the working model as incorrect? We propose leave-one-covariate-out or LOCO inference, proceeds as follows. Let $\widehat{\mu}$ denote our estimate of the mean function on (X_i, Y_i), $i \in \mathcal{I}_1$ for some $\mathcal{I}_1 \subseteq \{1, \cdots, n\}$. Also, let $\widehat{\mu}_{(-j)}$ denote the refitted mean function on $(X_i(-j), Y_i)$, $i \in \mathcal{I}_1$ where $X_i(-j) = (X_i(1), \cdots, X_i(j-1), X(j+1), \cdots, X_i(d))$. The excess prediction error of covariate j at a new i.i.d. draw (X_{n+1}, Y_{n+1}) is given by

$$\Delta_j(X_{n+1}, Y_{n+1}) = |Y_{n+1} - \widehat{\mu}_{(-j)}(X_{n+1})| - |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

6.1 Local Measure of Variable Importance

We can construct a valid prediction interval for the random variable $\Delta_j(X_{n+1}, Y_{n+1})$ using conformal prediction bands. Let C denote a conformal prediction set for Y_{n+1} given X_{n+1}, having coverage $1-\alpha$. Note that we use $\mathcal{I}_1 = \{1, \cdots, n\}$ to fit $\widehat{\mu}$ and $\widehat{\mu}_{(-j)}$ for the full method, and $\mathcal{I}_1 \subsetneq \{1, \cdots, n\}$ for the split method. Now define

$$W_j(x) = \{|y - \widehat{\mu}_{(-j)}(x)| - |y - \widehat{\mu}(x)| : y \in C(x)\}$$

From the finite-sample validity of C, we have

$$\mathbb{P}(\Delta_j(X_{n+1}, Y_{n+1}) \in W_j(X_{n+1}), \text{ for all } j = 1, \cdots, d) \geq 1 - \alpha$$

To summarize the effect of covariate j, we can plot the intervals $W_j(X_i)$ for $i = 1, \cdots, n$. By a low-dimenional additive model example, we show that the intervals above zero indicate that such covariates are important. Furthermore, the locations at which the intervals lie above zero corresponds each component.

6.2 Global Measures of Variable Importance

For a global measure of variable importance, we can focus on the distribution of $\Delta_j(X_{n+1}, Y_{n+1})$, marginally over (X_{n+1}, Y_{n+1}). We rely on a splitting approach and let $\mathcal{D}_k = \{(X_i, Y_i) : i \in \mathcal{I}_k\}$. Define

$$\theta_j = \mathbb{E}\left[\Delta_j(X_{n+1}, Y_{n+1}) \mid \mathcal{D}_1\right]$$

Then, we can obtain the asymptotic $1-\alpha$ confidence interval

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

where $\widehat{\theta}_j = (n/2)^{-1} \sum_{i \in \mathcal{I}_2} \Delta_j(X_i, Y_i)$ and s_j^2 is the analogous sample variance measured on $\mathcal{D}_2$.

It may be more stable to consider the median parameter

$$m_j = \text{median}\left[\Delta_j(X_{n+1}, Y_{n+1}) \mid \mathcal{D}_1\right]$$

We can conduct nonasymptotic inferences about m_j using standard, nonparametric tests such as the sign test or the Wilcoxon signed-rank test, applied to $\Delta_j(X_i, Y_i)$, $i \in \mathcal{I}_2$. By a high-dimensional linear regression problem example, we show that the inference performs well.

7 Conclusion

Conformal prediction bands are robust against model misspecification, as long as the data are i.i.d. The "out-of-sample-fitting" of conformal inference guarantees finite-sample coverage property compared to jackknife, bootstrap, and cross-validation. We can use the conformal prediction band as a diagnostic and comparison tool for regression function estimators.

Lastly, it would be practically appealing to develop novel methods that more efficiently combine results from multiple splits. Also, it would be interesting to see how conformal inference can help with model-free variable selection.

References

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). *Simultaneous analysis of lasso and dantzig selector.* The Annals of Statistics, p. 1705-1732.

Thakurta, A. G., & Smith, A. (2013). *Differentially private feature selection via stability arguments, and the robustness of the lasso.* In Conference on Learning Theory, p. 819-850.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World.* Springer.