Distribution free Prediction and Regression

Anirban Chatterjee

Department of Statistics and Data Science

02/22/2022

Penn
University of Pennsylvania

Outline

Table of Contents

Penn
UNIVERSITY of PENNSYLVANIA

Introduction

- Goal of conformal prediction: Without knowledge of underlying distribution, produce ”valid” bands using observed data.
- Goal of distribution free inference: Without knowledge of underlying distribution, ”infer” about some property of that distribution.

Introduction

- Goal of conformal prediction: Without knowledge of underlying distribution, produce "valid" bands using observed data.
- Goal of distribution free inference: Without knowledge of underlying distribution, "infer" about some property of that distribution.
  - Density level sets
  - Regression confidence bands (property of joint distribution)

A Refresher on Conformal Prediction

- Given $Z_{\text{Obs}} = \{Z_i : 1 \leq i \leq n\} \sim P$ (unknown) $\Longrightarrow C_n\left(Z_{\text{Obs}}\right)$ such that

$$\mathbb{P}_{Z_{\text{Obs}}, Z_{n+1}}\left(Z_{n+1} \in C_n\left(Z_{Obs}\right)\right) \geq 1 - \alpha \rightarrow C_n \text{ is valid }.$$

- Only requirement: Z_{Obs}, Z_{n+1} are *Exchangeable*.

Penn
UNIVERSITY *of* PENNSYLVANIA

A Refresher on Conformal Prediction

Algorithm:

- Non-Conformity Score: $\sigma_i = \sigma\left(\{Z_{\text{Obs}}, Z_{n+1}\}; Z_i\right) \leftarrow Z_i \not\sim \{Z_j : 1 \leq j \leq n+1\}$, where $\sigma \leftarrow$ is permutation invariant in first entry.

University of Pennsylvania

A Refresher on Conformal Prediction

Algorithm:

- Non-Conformity Score:
  $\sigma_i = \sigma\left(\{Z_{\text{Obs}}, Z_{n+1}\}; Z_i\right) \leftarrow Z_i \nsim \{Z_j : 1 \leq j \leq n+1\}$, where $\sigma \leftarrow$ is permutation invariant in first entry.
- *Prediction region*:

$$C_n\left(Z_{\text{Obs}}, \sigma\right) = \left\{ z : \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\left[\sigma_i(Z_{n+1} = z) \leq \sigma_{n+1}(Z_{n+1} = z)\right] \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \right\}$$

Table of Contents

Density Level sets

- $P \leftarrow$ Distribution and $p \leftarrow$ density.
- Where is most of the probability mass concentrated?

Density Level sets

- $P \leftarrow$ Distribution and $p \leftarrow$ density.
- Where is most of the probability mass concentrated?

$$L(t) := \left\{ y \in \mathbb{R}^d : p(y) \geq t \right\}$$

Density Level sets

- $P \leftarrow$ Distribution and $p \leftarrow$ density.
- Where is most of the probability mass concentrated?

$$L(t) := \left\{ y \in \mathbb{R}^d : p(y) \geq t \right\}$$

- $t(\alpha)$ = Lower α quantile of $p(Y)$, $Y \sim P$.
- The *minimum volume prediction* set is equivalent to *density level sets*.

$$\boldsymbol{C}(\boldsymbol{\alpha}) := L(t(\alpha)) = \arg\min_{\mathbb{C}} m(C),\ \mathbb{C} = \{C : P(C) \geq 1 - \alpha\}.$$

Penn
UNIVERSITY *of* PENNSYLVANIA

Distribution Free Prediction Sets

Goal: To find C_n based on observed data such that,

- C_n is valid.
- $m\left(C_n \triangle C(\alpha)\right) = o_{\mathbb{P}}(1)$.

Penn
UNIVERSITY of PENNSYLVANIA

Distribution Free Prediction Sets

Goal: To find C_n based on observed data such that,

- C_n is valid.
- $m\left(C_n \triangle C(\alpha)\right) = o_{\mathbb{P}}(1)$.
- Checking $y \in C_n$ = O(n)

Distribution Free Prediction Sets

Goal: To find C_n based on observed data such that,

- C_n is valid.
- $m\left(C_n \triangle C(\alpha)\right) = o_{\mathbb{P}}(1)$.
- Checking $y \in C_n$ = O(n)

Main Idea:

- Use Conformal Prediction with a particular non-conformity score $\rightarrow$ Kernel Density Estimator

Distribution Free Prediction Sets

Goal: To find C_n based on observed data such that,

- C_n is valid.
- $m\left(C_n \triangle C(\alpha)\right) = o_{\mathbb{P}}(1)$.
- Checking $y \in C_n$ = O(n)

Main Idea:

- Use Conformal Prediction with a particular non-conformity score $\rightarrow$ Kernel Density Estimator
- Sandwiching,

Kernel Level set $\subseteq$ Conformal Prediction set $\subseteq$ Kernel Level set

Kernel Density Non-Conformity Score

Kernel Density Estimator: Given $\{Z_i : 1 \leq i \leq n\}$,

$$\widehat{p}_n(u) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{u - Z_i}{h}\right)$$

Kernel Density Non-Conformity Score: Considering $Z_{n+1} = z$ and $\widehat{p}_n^z(u) = \widehat{p}_{n+1}(u)$,

$$\sigma_i = \frac{1}{\widehat{p}_n^z(Z_i)} \text{ for all } 1 \leq i \leq n+1.$$

Kernel Density Prediction set $\left(\widehat{C}_n(\alpha)\right)$

$$z : \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left[\frac{1}{\widehat{p}_n^z(Z_i)} \leq \frac{1}{\widehat{p}_n^z(z)}\right] \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}$$

Equivalently

$$z : \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left[\widehat{p}_n^z(Z_i) \leq \widehat{p}_n^z(z)\right] \geq \frac{\lfloor (n+1)\alpha \rfloor}{n+1} = \widetilde{\alpha}$$

Kernel Density Prediction set $\left(\widehat{C}_n(\alpha)\right)$

$$z : \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left[\frac{1}{\widehat{p}_n^z(Z_i)} \leq \frac{1}{\widehat{p}_n^z(z)}\right] \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}$$

Equivalently

$$z : \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left[\widehat{p}_n^z(Z_i) \leq \widehat{p}_n^z(z)\right] \geq \frac{\lfloor (n+1)\alpha \rfloor}{n+1} = \widetilde{\alpha}$$

Valid Interval

Penn
UNIVERSITY of PENNSYLVANIA

Sandwiching - An Approximation

- $\widehat{p}_n \leftarrow$ Kernel density estimator, based on $\{Z_i : 1 \leq i \leq n\}$.
- $L_n(t) = \{z : \widehat{p}_n(z) \geq t\} \leftarrow$ Level sets of Kernel density estimator.
- **Rank** $\{Z_i : 1 \leq i \leq n\}$ according to $\{\widehat{p}_n(Z_i) : 1 \leq i \leq n\}$.

Sandwiching - An Approximation

- $\widehat{p}_n \leftarrow$ Kernel density estimator, based on $\{Z_i : 1 \leq i \leq n\}$.
- $L_n(t) = \{z : \widehat{p}_n(z) \geq t\} \leftarrow$ Level sets of Kernel density estimator.
- **Rank** $\{Z_i : 1 \leq i \leq n\}$ according to $\{\widehat{p}_n(Z_i) : 1 \leq i \leq n\}$.

Lemma

When $K(0) = \sup K(u)$ *then,*

$$\boldsymbol{L_n^-} := L_n\left(\widehat{p}\left(Z_{(\lfloor (n+1)\alpha \rfloor)}\right)\right) \subseteq \widehat{C}_n(\alpha)$$

$$\subseteq \boldsymbol{L_n^+} := L_n\left(\widehat{p}\left(Z_{(\lfloor (n+1)\alpha \rfloor)}\right) - O\left(\frac{h^d}{n}\right)\right)$$

Penn UNIVERSITY of PENNSYLVANIA

Sandwiching - An Approximation

- L_n^+ is **valid**
- Checking $y \in L_n^+ = O(n)$.

Sandwiching: Proof Sketch

For $i \leq \lfloor (n+1)\alpha \rfloor$ and $z \in L_n\left(\widehat{p}\left(Z_{(\lfloor (n+1)\alpha \rfloor)}\right)\right)$,

$$\widehat{p}_n^z(z) - \widehat{p}_n^z\left(Z_{(i)}\right) \geq c_n\left(\widehat{p}_n(z) - \widehat{p}_n\left(Z_{(i)}\right)\right) \geq 0.$$

Implies

$$\frac{1}{n+1}\left(\sum_{i=1}^{n} [\widehat{p}_n^z(Z_i) \leq \widehat{p}_n^z(z)] + 1\right) \geq \frac{\lfloor (n+1)\alpha \rfloor}{n+1} \implies z \in \widehat{C}_n(\alpha)$$

The upper bound follows similarly by considering $y \notin L_n^+$ and $i \geq \lfloor (n+1)\alpha \rfloor$.

Penn
UNIVERSITY *of* PENNSYLVANIA

Kernel Density Prediction: Accuracy

Goal:

$$m\left(\widehat{C}\triangle C(\alpha)\right) = o_{\mathbb{P}}(1),\ \widehat{C} \in \left\{L_n^-, \widehat{C}_n(\alpha), L_n^+\right\}$$

Technical Assumptions:

- $p \leftarrow \beta$ Hölder smooth, $K \leftarrow$ order β
- Distribution function of $p(Y)$, $Y \sim P$ is well behaved near $t(\alpha)$,

$$c_1|\epsilon|^\gamma \leq |P\left(p(Y) \leq t(\alpha) + \epsilon\right) - \alpha| \leq c_2|\epsilon|^\gamma$$

$$\uparrow$$

γ-exponent condition.

Kernel Density Prediction: Accuracy

Theorem

If $h \approx \left(\frac{\log n}{n}\right)^{c_{p,d}}$ *then,*

$$m\left(\widehat{C} \triangle C(\alpha)\right) = o_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{c_{p,\alpha}}\right)$$

Accuracy: Proof Sketch I

Consider $t_n = \widehat{p}\left(Z_{(\lfloor(n+1)\alpha\rfloor)}\right)$. Define $R_n = \|\widehat{p}_n - p\|_\infty$, and

$V_n = \sup_{t>0}\left|P_n\left(L^l(t)\right) - P\left(L^l(t)\right)\right|$ where $L^l(t) = \{y : p(y) \leq t\}$

Lemma

$$|t_n - t(\alpha)| = O_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{b_{p,\alpha}}\right)$$

Proof:

Accuracy: Proof Sketch II

Consider $\alpha_n = \dfrac{\lfloor (n+1)\alpha \rfloor}{n}$. Then using γ-exponent condition,

$$|t(\alpha_n) - t(\alpha)| = O\left(n^{-1/\gamma}\right) \tag{1}$$

Considering G and G_n to be distribution corresponding to p and $\widehat{p}_n$. Using definition of L^l, it can be easily observed that,

$$G\left(t - R_n\right) - V_n \leq G_n(t) \leq G(t + R_n) + V_n \tag{2}$$

Penn
University of Pennsylvania

Accuracy: Proof Sketch III

Using standard empirical process theory,

$$V_n = O_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2}}\right)$$

and

$$R_n = O_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{a_{p,\alpha}}\right)$$

Accuracy: Proof Sketch IV

Consider $W_n = R_n + (2V_n/c_1)^{1/\gamma}$. Then for large enough n, using (2),

$$G_n\left(t(\alpha_n) - W_n\right) < \alpha_n < G_n\left(t(\alpha_n) + W_n\right)$$

Implying $|t_n - t(\alpha_n)| \leq W_n$, and then using bounds on W_n in combination with (1) completes the proof. $\square$

$$\begin{aligned} L_n^- \triangle C(\alpha) &= \{\widehat{p}_n \geq t_n, p < t(\alpha)\} \cup \{\widehat{p}_n < t_n, p \geq t(\alpha)\} \\ &\subseteq \{t(\alpha) - |t_n - t(\alpha)| - R_n \leq p < t(\alpha)\} \cup \\ &\quad \{t(\alpha) \leq p \leq t(\alpha) + |t_n - t(\alpha)| + R_n\} \end{aligned} \tag{3}$$

Accuracy: Proof Sketch V

Observe that on $L_n^- \triangle C(\alpha)$,

$$p \geq t(\alpha) - |t_n - t(\alpha)| - R_n$$

and hence,

$$\left(t(\alpha) - |t(\alpha) - t_n| - R_n\right) m \left(L_n^- \triangle C(\alpha)\right) \leq P \left(L_n^- \triangle C(\alpha)\right) \tag{4}$$

Accuracy: Proof Sketch VI

For large enough n using above lemma, (4) becomes,

$$m\left(L_n^- \triangle C(\alpha)\right) \leq \frac{P\left(L_n^- \triangle C(\alpha)\right)}{\left(t(\alpha) - \left|t(\alpha) - t_n\right| - R_n\right)}$$

and finally using the expansion of $L_n^- \triangle C(\alpha)$ from (3),

$$m\left(L_n^- \triangle C(\alpha)\right) \leq \frac{c}{t(\alpha)}\left(O\left(\left(\frac{\log n}{n}\right)^{b_{p,\alpha}} + \left(\frac{\log n}{n}\right)^{a_{p,\alpha}}\right)\right)^{\gamma} \text{ w.h.p.}$$

Bandwidth Selection

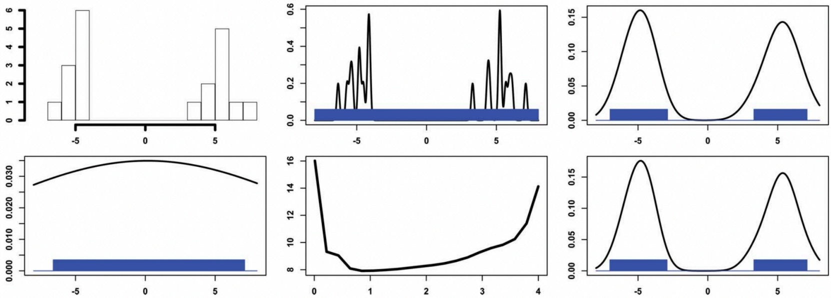


Figure 1: Bandwidth and Conformal Prediction set

Bandwidth Selection

Lemma

$$\mathbb{E}m\left(\widehat{C}\triangle C(\alpha)\right) \leq c\left[\mathbb{E}(m(\widehat{C}) + c_0)\right]^{1/2}$$

Bandwidth Selection

Lemma

$$\mathbb{E}m\left(\widehat{C}\triangle C(\alpha)\right) \leq c\left[\mathbb{E}(m(\widehat{C})+c_0)\right]^{1/2}$$

Choose bandwidth to minimize width of prediction set

Table of Contents

Penn
University of Pennsylvania

Predictive Inference for Regression

Problem Setup:

- $\{Z_i = (Y_i, X_i) : 1 \leq i \leq n\} \overset{i.i.d}{\sim} P$
- $\mu(x) = \mathbb{E}\left(Y|X = x\right)$ is the regression function.
- No assumptions on P or μ.

Objective:

For a new feature value X_{n+1},

Produce $C_n = C_n(\{Z_1, \cdots, Z_n\}, X_{n+1}) \to \mathbb{P}\left(Y_{n+1} \in C_n\right) \geq 1 - \alpha$

Penn
UNIVERSITY of PENNSYLVANIA

Conformal Prediction for Regression

$\widehat{\mu} \leftarrow$ symmetric regression estimator

Non-Conformity Scores

- Augmented Data $\leftarrow \{Z_i = (Y_i, X_i) : 1 \leq i \leq n\} \cup \{(y, X_{n+1})\}$
- $\widehat{\mu}_y \leftarrow$ Augmented data estimator.

$$\sigma_i = R_i(y) = |Y_i - \widehat{\mu}_y(X_i)|, 1 \leq i \leq n; \sigma_{n+1} = R_{n+1}(y) = |y - X_{n+1}|$$

Conformal Prediction for Regression

Prediction region for regression: $\left(C_n(X_{n+1})\right)$

$$y : \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbf{1}\left[R_i(y) \leq R_{n+1}(y)\right] \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}$$

Penn
UNIVERSITY of PENNSYLVANIA

Conformal Prediction for Regression

Validity of Prediction Region

$$1 - \alpha \leq \mathbb{P}\left(Y_{n+1} \in C_n(X_{n+1})\right) \leq 1 - \alpha + \frac{1}{n+1}$$

Split-Conformal Prediction for Regression

Full Conformal $\leftarrow$ Computationally Intensive.

$\leftarrow$ Requires retraining.

Penn
UNIVERSITY of PENNSYLVANIA

Split-Conformal Prediction for Regression

Algorithm 2 Split Conformal Prediction

Input: Data (X_i, Y_i), $i = 1, \dots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$
Output: Prediction band, over $x \in \mathbb{R}^d$
Randomly split $\{1, \dots, n\}$ 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 $x \in \mathbb{R}^d$

Figure 2: Split Conformal Prediction for Regression

Split-Conformal Prediction for Regression

Validity of Prediction Region

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

Split-Conformal Prediction for Regression

Validity of Prediction Region

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

In-Sample Coverage Guarantee

$$\frac{2}{n} \sum_{i \in \mathcal{I}_2} \mathbf{1}\left[Y_i \in C_{\text{split}}(X_i)\right] \approx 1 - \alpha \text{ w.h.p.}$$

Accuracy of Conformal Prediction for Regression

- Length of Conformal Interval $\approx$ Length of ”Oracle” Interval w.h.p.

Accuracy of Conformal Prediction for Regression

- Length of Conformal Interval $\approx$ Length of "Oracle" Interval w.h.p.
- $m\left(\text{Confomal Interval}\triangle\text{"Oracle" Interval}\right) = o_{\mathbb{P}}(1)$

Accuracy of Conformal Prediction for Regression

- Length of Conformal Interval $\approx$ Length of "Oracle" Interval w.h.p.
- $m\left(\text{Confomal Interval} \triangle \text{"Oracle" Interval}\right) = o_{\mathbb{P}}(1)$

Technical Assumptions

- I.I.D. data
- Noise $= \epsilon = Y - \mu(X)$ has a non-increasing density symmetric around 0.

Oracle Prediction Bands

Super Oracle

- Knows everything

$$C_s(x) = [\mu(x) - q_\alpha, \mu(x) + q_\alpha], \; q_\alpha \leftarrow \text{ Upper } \alpha \text{ quantile of } |\epsilon|$$

"Regular" Oracle

- Knows distribution of $Y - \widehat{\mu}_n(X)$, where $(X, Y) \sim P$.

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

$$q_{n,\alpha} \leftarrow \text{ Upper } \alpha \text{ quantile of } |Y - \widehat{\mu}_n(X)|$$

Comparing the Oracles

Theorem

- $F, f \leftarrow$ *distribution, density of* $|\epsilon|$
- $F_n, f_n \leftarrow$ *distribution, density of* $|Y - \widehat{\mu}_n(X)|$.

$$\|F_n - F\|_\infty \leq c_f \, \mathbb{E}\left[\widehat{\mu}_n(X) - \mu(X)\right]^2$$

- *Under regularity conditions on* f *near* q_α,

$$|q_{n,\alpha} - q_\alpha| \leq b_f \, \mathbb{E}\left[\widehat{\mu}_n(X) - \mu(X)\right]^2$$

Approximating the "Regular" Oracle

Split Conformal

Theorem

If $\|\widehat{\mu}_n - \mu\|_\infty = o_{\mathbb{P}}(1)$, *then*

$$\textit{Length}\left(C_{split}\right) - 2q_{n,\alpha} = o_{\mathbb{P}}(1)$$

Approximating the “Regular” Oracle

Full Conformal

Theorem

- $Y \in \mathcal{Y} \leftarrow$ *a compact interval.*
- $\widehat{\mu}_{n,(X,y)} \leftarrow$ *fitted regression function using augmented data* $(X_{n+1} = X, Y_{n+1} = y)$.
- $\sup_{y \in \mathcal{Y}} \|\widehat{\mu}_n - \widehat{\mu}_{n,(X,y)}\|_\infty = o_{\mathbb{P}}(1)$, *then*

$$\textit{Length}\left(C_n(X)\right) - 2q_{n,\alpha} = o_{\mathbb{P}}(1)$$

Approximating the Super Oracle

Theorem

- *Same assumptions as before.*
- $\mathbb{E}\left[\widehat{\mu}_n(X) - \mu(X)\right]^2 = o(1)$

$$m\left(\widehat{C} \triangle C_s\right) = o_{\mathbb{P}}(1),\ C \in \{C_n, C_{split}\}$$

Table of Contents

Extension of Conformal Inference

- **In-Sample Split Conformal Inference**
- **Model-Free Variable Importance**

In-Sample Split Conformal Inference

Problem:

Want C_n based on samples $\{(X_i, Y_i) : 1 \leq i \leq n\}$ such that,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[Y_i \in C_n(X_i)\right] \approx 1 - \alpha$$

Penn
UNIVERSITY of PENNSYLVANIA

In-Sample Split Conformal Inference

Problem:

Want C_n based on samples $\{(X_i, Y_i) : 1 \le i \le n\}$ such that,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[Y_i \in C_n(X_i)\right] \approx 1 - \alpha$$

A Simple Solution: $C_n(X_i) \leftarrow$ using $\{Z_j = (X_j, Y_j) : j \neq i\}$.

In-Sample Split Conformal Inference

Problem:

Want C_n based on samples $\{(X_i, Y_i) : 1 \leq i \leq n\}$ such that,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[Y_i \in C_n(X_i)\right] \approx 1 - \alpha$$

A Simple Solution: $C_n(X_i) \leftarrow$ using $\{Z_j = (X_j, Y_j) : j \neq i\}$.

$\uparrow$

- Computationally Intensive $\leftarrow$ multiplies by $O(n)$.
- Complex dependency structure, analytically intractable.

Penn
UNIVERSITY of PENNSYLVANIA

Rank-One-Out Split Conformal

$$\{(X_i, Y_i) : 1 \le i \le n\} \to \{(X_i, Y_i) : i \in \mathcal{I}_1\} \sqcup \{(X_i, Y_i) : i \in \mathcal{I}_2\}$$

$$\downarrow$$

$$k \in \{1, 2\}, \quad \widehat{\mu}_k \leftarrow \mathcal{I}_k$$

$$\downarrow$$

$$i \notin \mathcal{I}_k \to R_i = |Y_i - \widehat{\mu}_k(X_i)|$$

$$\downarrow$$

$$C_{\text{roo}}(X_i) = [\widehat{\mu}_k(X_i) - d_i, \widehat{\mu}_k(X_i) + d_i], \; d_i = q_{1-\alpha}\left(R_j : j \notin \mathcal{I}_k, j \neq i\right)$$

Penn
UNIVERSITY of PENNSYLVANIA

Rank-One-Out Split Conformal

Theorem

$$1 - \alpha \preceq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[Y_i \in C_{roo}(X_i)\right] \preceq 1 - \alpha + \frac{2}{n} \quad \textit{w.h.p.}$$

Model Free Variable Importance

Q. How to measure of each covariate in a prediction model?

Model Free Variable Importance

Q. How to measure of each covariate in a prediction model?

In linear model $\leftarrow$ Estimated Coefficients

Model Free Variable Importance

Q. How to measure of each covariate in a prediction model?

In linear model $\longleftarrow$ Estimated Coefficients

Model-Free general method $\longrightarrow$ **Leave-One-Covariate-Out** (LOCO).

LOCO

Importance for covariate j

$$\{(X_i, Y_i) : 1 \leq i \leq n\} \to \widehat{\mu}$$

$$X_i(-j) = (X_i(1), \cdots, X_i(j-1), X_i(j+1), \cdots, X_i(d))$$

$$\{(X_i(-j), Y_i) : 1 \leq i \leq n\} \to \widehat{\mu}_{(-j)}$$

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

$\uparrow$

Excess Prediction Error

LOCO

Importance for covariate j

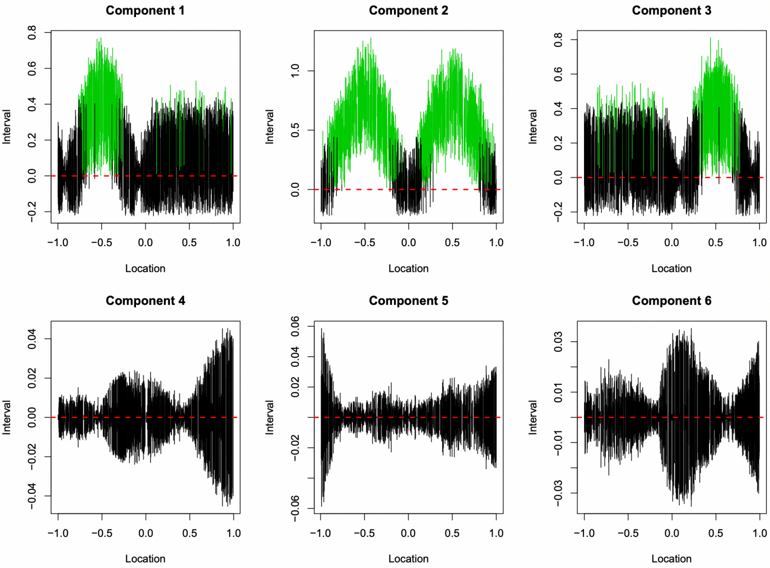
$$W_j(x) = \{\Delta_j(x, y) : y \in C_n(x)\}$$

$$W_j(X_i) \iff \text{Variable Importance}$$

An Example:

$\mu(x) = \sum_{j=1}^{6} f_j(x_j)$, where $f_4 = f_5 = f_6 = 0$. $X_i \overset{i.i.d}{\sim} \text{Unif}[-1, 1]^d$,

$$Y = \mu(X) + \mathbf{N}(0, 1)$$


Component 1
Component 2
Component 3
Component 4
Component 5
Component 6
Interval
Location

References I

- Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman, *Distribution-free predictive inference for regression*, J. Amer. Statist. Assoc. **113** (2018), no. 523, 1094–1111. MR 3862342
- Jing Lei, James Robins, and Larry Wasserman, *Distribution-free prediction sets*, J. Amer. Statist. Assoc. **108** (2013), no. 501, 278–287. MR 3174619