Lecture 2
Distribution-free inference: limits & open questions

Rina Foygel Barber

http://www.stat.uchicago.edu/~rina/

Limitations of distribution-free prediction

The guarantee for conformal prediction / holdout methods:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

↗

w.r.t. distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ i.i.d. from any distribution

The drawbacks:

- The guarantee is *on average* over the training data
- The guarantee is *on average* over the test point X_{n+1}
- And, what if the data is not independent or not identically distrib.?

---

[1] Vovk 2012, *Conditional validity of inductive conformal predictors*

[2] Work in progress, Bian & B. 2021

Limitations of distribution-free prediction

The guarantee for conformal prediction / holdout methods:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

↗

w.r.t. distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ i.i.d. from any distribution

The drawbacks:

- The guarantee is *on average* over the training data

  For holdout, concentration ensures the coverage holds w.h.p. over training data[1]

  For full conformal / jackknife+, no guarantee is possible![2]

- The guarantee is *on average* over the test point X_{n+1}

- And, what if the data is not independent or not identically distrib.?

---

[1] Vovk 2012, *Conditional validity of inductive conformal predictors*

[2] Work in progress, Bian & B. 2021

Limitations of distribution-free prediction

The guarantee for conformal prediction / holdout methods:

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right\} \geq 1 - \alpha$$

↗

w.r.t. distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ i.i.d. from any distribution

The drawbacks:

- The guarantee is *on average* over the training data

  For holdout, concentration ensures the coverage holds w.h.p. over training data[1]

  For full conformal / jackknife+, no guarantee is possible![2]

- The guarantee is *on average* over the test point X_{n+1}

  Will discuss this next

- And, what if the data is not independent or not identically distrib.?

[1] Vovk 2012, *Conditional validity of inductive conformal predictors*

[2] Work in progress, Bian & B. 2021

Limitations of distribution-free prediction

The guarantee for conformal prediction / holdout methods:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

↗

w.r.t. distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ i.i.d. from any distribution

The drawbacks:

- The guarantee is *on average* over the training data
  For holdout, concentration ensures the coverage holds w.h.p. over training data[1]
  For full conformal / jackknife+, no guarantee is possible![2]
- The guarantee is *on average* over the test point X_{n+1}
  Will discuss this next
- And, what if the data is not independent or not identically distrib.?
  Discussed in lecture 1 — covariate shift, time series, ... — need assumptions!

[1] Vovk 2012, *Conditional validity of inductive conformal predictors*

[2] Work in progress, Bian & B. 2021

Conditional prediction

Is it possible to provide prediction that's valid conditional on X_{n+1}, i.e.,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1}\right\} \geq 1 - \alpha \text{ ?}$$

(Motivation—the marginal guarantee doesn't exclude, e.g.,

90% of individuals have 100% coverage / 10% of individuals have 0% coverage)

[3] Vovk 2012, *Conditional validity of inductive conformal predictors*

Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Conditional prediction

Is it possible to provide prediction that's valid conditional on X_{n+1}, i.e.,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1}\right\} \geq 1 - \alpha \text{ ?}$$

(Motivation—the marginal guarantee doesn't exclude, e.g.,

90% of individuals have 100% coverage / 10% of individuals have 0% coverage)

- If X is nonatomic (i.e., $P_X(x) = 0$ for all $x \in \mathcal{X}$), impossible—
  $\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right] = \infty$ for any $\widehat{C}_n$ that's valid distribution-free[3]

[3] Vovk 2012, *Conditional validity of inductive conformal predictors*
Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Conditional prediction

Is it possible to provide prediction that's valid conditional on X_{n+1}, i.e.,

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1} \right\} \geq 1 - \alpha \text{ ?}$$

(Motivation—the marginal guarantee doesn't exclude, e.g.,

90% of individuals have 100% coverage / 10% of individuals have 0% coverage)

- If X is nonatomic (i.e., $P_X(x) = 0$ for all $x \in \mathcal{X}$), impossible—

$$\underbrace{\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right]}_{\text{expected length when data } \overset{\text{iid}}{\sim} P} = \infty \text{ for any } \underbrace{\widehat{C}_n \text{ that's valid distribution-free}}_{\text{coverage must hold when data } \overset{\text{iid}}{\sim} \text{ any distribution}}$$ [3]

[3] Vovk 2012, *Conditional validity of inductive conformal predictors*
Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Conditional prediction

Can we relax the notion of conditionally valid coverage, to obtain a nontrivial $\widehat{C}_n$?

$(1-\alpha, \delta)$-conditional coverage:[4] for any P & any $\mathcal{X}_*$ with $P_X(\mathcal{X}_*) \geq \delta$,

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1} \in \mathcal{X}_* \right\} \geq 1-\alpha \text{ w.r.t. data} \overset{\text{iid}}{\sim} P.$$

[4] B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*

Conditional prediction

Can we relax the notion of conditionally valid coverage, to obtain a nontrivial $\widehat{C}_n$?

$(1-\alpha,\delta)$-conditional coverage:[4] for any P & any $\mathcal{X}_*$ with $P_X(\mathcal{X}_*) \geq \delta$,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1} \in \mathcal{X}_*\right\} \geq 1-\alpha \text{ w.r.t. data} \overset{\text{iid}}{\sim} P.$$

A trivial solution: any method with $(1-\alpha\delta)$-marginal coverage, automatically satisfies $(1-\alpha,\delta)$-conditional coverage

- The problem — any interval w/ $(1-\alpha\delta)$ coverage will be very wide

[4] B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*

Conditional prediction

Theorem: for nonatomic P_X, the trivial solution is essentially optimal:
If $\widehat{C}_n$ satisfies $(1-\alpha,\delta)$-CC, then

$$\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right] \geq \left(\begin{array}{c} \text{min. length of any oracle method} \\ \text{with } 1-\alpha\delta \text{ coverage for } P \end{array} \right)$$

Conditional prediction

Conditional on bins: partition $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$,
& require $\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \;\middle|\; X_{n+1} \in \mathcal{X}_k \right\} \geq 1 - \alpha$ for each k[5]

- For each k, data points $\{(X_i, Y_i) : X_i \in \mathcal{X}_k\}$ are exchangeable $\rightsquigarrow$ run CP separately for each k to guarantee bin-conditional coverage
- Note — the model $\widehat{\mu}$ can still be fitted on the entire data set!

An application — fairness with respect to subpopulations[6]

[5] Vovk 2012, *Conditional validity of inductive conformal predictors*
Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*
B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*
[6] Romano, B., Sabatti, Candès 2019, *With malice toward none: assessing uncertainty via equalized coverage*

Conditional prediction

Extensions:

Combining distribution-free inference with assumption-based inference:[7]

- Estimate the conditional distribution of $Y|X \rightsquigarrow \widehat{F}(y|x)$
- Use nonconformity score is $S(x,y) = |\widehat{F}(y|x) - 0.5|$
  - CP is valid with any score $\Rightarrow$ marginal coverage
  - If $\widehat{F}$ satisfies consistency conditions $\Rightarrow$ conditional coverage

[7] Chernozhukov et al 2019, *Distributional conformal prediction*

Conditional prediction

Extensions:

A localized form of the prediction guarantee—[8]

construct the PI using a kernel around the test point,

e.g., only the nearest neighbors of the test point

$\rightarrow$ achieves a local version of predictive validity

[8] Guan 2020, *Conformal prediction with localization*

Inference for regression

What about inference for regression (confidence not prediction)?

Define marginal validity for confidence intervals:[9]

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

w.r.t. data $\stackrel{\text{iid}}{\sim} P$ for any distribution P, where $\mu_P(x) = \mathbb{E}\left[Y \mid X = x\right]$

[9] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[10] B. 2020, *Is distribution-free inference possible for binary regression?*

Inference for regression

What about inference for regression (confidence not prediction)?

Define marginal validity for confidence intervals:[9]

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

w.r.t. data $\stackrel{\text{iid}}{\sim} P$ for any distribution P, where $\mu_P(x) = \mathbb{E}\left[Y \mid X = x\right]$

Special case: binary regression $\rightsquigarrow \mu_P(x) = \mathbb{P}\left\{Y = 1 \mid X = x\right\}$

[9] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[10] B. 2020, *Is distribution-free inference possible for binary regression?*

What about inference for regression (confidence not prediction)?

Define marginal validity for confidence intervals:[9]

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha$$

w.r.t. data $\stackrel{\text{iid}}{\sim} P$ for any distribution P, where $\mu_P(x) = \mathbb{E}\left[Y \mid X = x\right]$

Special case: binary regression $\rightsquigarrow \mu_P(x) = \mathbb{P}\left\{Y = 1 \mid X = x\right\}$

Theorem:[10] If X is nonatomic, then

$$\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right] \geq \underbrace{\text{constant lower bound}}_{\text{depends on } P \text{ and } \alpha \text{ but not on } n}$$

Example: if $Y|X \sim \text{Bernoulli}(0.5)$, $\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right] \geq 1-\alpha$

(compare to trivial solution: $\widehat{C}_n(x) = [0,1]$ w.p. $1-\alpha$ or $\varnothing$ otherwise)

[9] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[10] B. 2020, *Is distribution-free inference possible for binary regression?*

Inference for regression

Intuition for why any distrib.-free conf. int. $\widehat{C}_n$ for μ_P must be wide...

Theorem:[11] If X is nonatomic, then any valid confidence interval $\widehat{C}_n$ is also a valid prediction interval:

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \; \forall \, P \quad \Rightarrow \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \; \forall \, P \text{ w/ } P_X \text{ nonatomic}$$

[11] B. 2020, *Is distribution-free inference possible for binary regression?*

[12] Medarametla & Candès 2021, *Distribution-free conditional median inference*

Inference for regression

Intuition for why any distrib.-free conf. int. $\widehat{C}_n$ for μ_P must be wide...

Theorem:[11] If X is nonatomic, then any valid confidence interval $\widehat{C}_n$ is also a valid prediction interval:

$$\mathbb{P}\left\{\mu_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \;\forall\, P \quad \Rightarrow \mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha \;\forall\, P \text{ w/ } P_X \text{ nonatomic}$$

A related result —

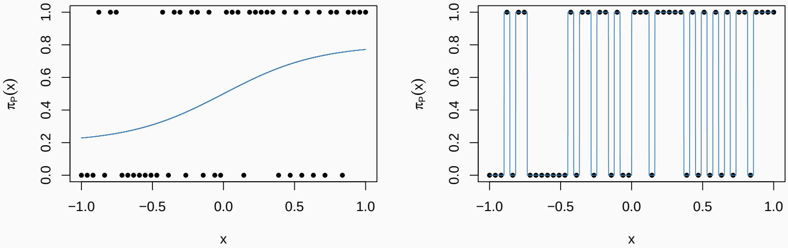
the same holds for any $\widehat{C}_n$ that covers the conditional median of $Y|X$[12]

[11]B. 2020, *Is distribution-free inference possible for binary regression?*

[12]Medarametla & Candès 2021, *Distribution-free conditional median inference*

Inference for regression

Intuition:

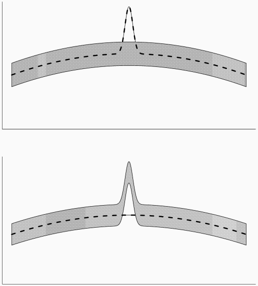


This challenge is related to the nonparametric regression literature — it is impossible to be adaptive to the level of smoothness[13]

[13] Giné & Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*

Inference for regression

From the nonparametric regression literature—[14]



"conservative failure" vs "liberal failure"

(figure from Genovese & Wasserman 2008)

Proposal: consider coverage of a surrogate function $\underbrace{\in \mathcal{F}}_{\text{functions } \tilde{f} \approx f \text{ that are smoother than } f}$ instead of true f

[14]Genovese & Wasserman 2008, *Adaptive confidence bands*

Calibration

Relaxing the goal of coverage of $\mu_P \rightsquigarrow$ calibration

- Perfect calibration: $\mathbb{E}[Y \mid f(X)] = f(X)$ almost surely

[15]Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Relaxing the goal of coverage of $\mu_P \rightsquigarrow$ calibration

- Perfect calibration: $\mathbb{E}[Y \mid f(X)] = f(X)$ almost surely
- Approx. calibration: $\left|\mathbb{E}[Y \mid f(X)] - f(X)\right| \leq \epsilon$ w.p.$\geq 1 - \alpha$

[15] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Relaxing the goal of coverage of $\mu_P \rightsquigarrow$ calibration

- Perfect calibration: $\mathbb{E}\left[Y \mid f(X)\right] = f(X)$ almost surely
- Approx. calibration: $\left|\mathbb{E}\left[Y \mid f(X)\right] - f(X)\right| \leq \epsilon$ w.p.$\geq 1 - \alpha$

Theorem:[15]

- Approx. calibration & d.f. inference for regression are equivalent — $\widehat{C}_n(X_{n+1}) = f(X_{n+1}) \pm \epsilon$ is a d.f. confidence interval for $\mu_P(X_{n+1})$

[15] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Relaxing the goal of coverage of $\mu_P \rightsquigarrow$ calibration

- Perfect calibration: $\mathbb{E}[Y \mid f(X)] = f(X)$ almost surely
- Approx. calibration: $\left|\mathbb{E}[Y \mid f(X)] - f(X)\right| \leq \epsilon$ w.p.$\geq 1 - \alpha$

Theorem:[15]

- Approx. calibration & d.f. inference for regression are equivalent —
  $\widehat{C}_n(X_{n+1}) = f(X_{n+1}) \pm \epsilon$ is a d.f. confidence interval for $\mu_P(X_{n+1})$

- Approx. calibration & d.f. prediction are equiv. for nonatomic P_X —
  $\widehat{C}_n(X_{n+1}) = f(X_{n+1}) \pm \epsilon$ is a d.f. prediction int. if P_X nonatomic

[15] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Calibration possible only if set of output values is $\leq$ countably infinite:[16]

- Let error level α be fixed, and let sample size $n \to \infty$
- A sequence of functions f_n is asymptotically calibrated if $\epsilon_n = o_P(1)$
- If there exists an asymptotically calibrated sequence f_n, then

$$\limsup_{n \to \infty} \left| \{\text{possible values of } f_n(X)\} \right| \leq \text{countably infinite}$$

[16] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Calibration possible only if set of output values is $\leq$ countably infinite:[16]

- Let error level α be fixed, and let sample size $n \to \infty$
- A sequence of functions f_n is asymptotically calibrated if $\epsilon_n = o_P(1)$
- If there exists an asymptotically calibrated sequence f_n, then

$$\limsup_{n\to\infty} \left|\{\text{possible values of } f_n(X)\}\right| \leq \text{countably infinite}$$

Intuitively, this connects to impossibility for regression —
$\{(f(X_i), Y_i)\}$ is a new regression problem $\rightsquigarrow$ impossible if $f(X)$ is nonatomic

[16] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

If $f(X)$ takes finitely many values... an example procedure:

- Use data $i = 1, \ldots, \frac{n}{2}$ to train $\widehat{\mu}(x)$, & partition into $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$
  (e.g., $\mathcal{X}_k = \{x : \text{cutoff}_{k-1} < \widehat{\mu}(x) \leq \text{cutoff}_k\}$)
- Use holdout set $i = \frac{n}{2} + 1, \ldots, n$ to estimate $\mathbb{E}[Y \mid X \in \mathcal{X}_k]$

[17] Gupta & Ramdas 2021, *Distribution-free calibration guarantees for histogram binning without sample splitting*

Calibration

If $f(X)$ takes finitely many values... an example procedure:

- Use data $i = 1, \ldots, \frac{n}{2}$ to train $\widehat{\mu}(x)$, & partition into $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$
  (e.g., $\mathcal{X}_k = \{x : \text{cutoff}_{k-1} < \widehat{\mu}(x) \leq \text{cutoff}_k\}$)
- Use holdout set $i = \frac{n}{2} + 1, \ldots, n$ to estimate $\mathbb{E}[Y \mid X \in \mathcal{X}_k]$

Binning can be data dependent without loss of validity —
we can use the holdout data to define the cutoffs between bins![17]
$\rightsquigarrow$ reduces loss of accuracy due to sample splitting
(still need to train $\widehat{\mu}$ separately)

[17] Gupta & Ramdas 2021, *Distribution-free calibration guarantees for histogram binning without sample splitting*

Beyond nonatomic

Returning to inference for regression (conf. int. for $\mathbb{E}[Y \mid X]$)...

Suppose P_X is instead *discrete*.
At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  ($P_X(x) \asymp 1$)
- Easy — each possible X value is observed many times
  ($P_X(x) \gg n^{-1}$)
- Medium — some X's are repeated, but most are unique
  ($n^{-2} \ll P_X(x) \ll n^{-1}$)
- Hard — w.h.p. the data set has no repeated X's ($P_X(x) \ll n^{-2}$)

Beyond nonatomic

Returning to inference for regression (conf. int. for $\mathbb{E}[Y \mid X]$)...

Suppose P_X is instead *discrete*.
At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  ($P_X(x) \asymp 1$) ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  ($P_X(x) \gg n^{-1}$)
- Medium — some X's are repeated, but most are unique
  ($n^{-2} \ll P_X(x) \ll n^{-1}$)
- Hard — w.h.p. the data set has no repeated X's ($P_X(x) \ll n^{-2}$)

Beyond nonatomic

Returning to inference for regression (conf. int. for $\mathbb{E}[Y \mid X]$)...

Suppose P_X is instead *discrete*.
At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  ($P_X(x) \asymp 1$) ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  ($P_X(x) \gg n^{-1}$) ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  ($n^{-2} \ll P_X(x) \ll n^{-1}$)
- Hard — w.h.p. the data set has no repeated X's ($P_X(x) \ll n^{-2}$)

Beyond nonatomic

Returning to inference for regression (conf. int. for $\mathbb{E}[Y \mid X]$)...

Suppose P_X is instead *discrete*.
At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  ($P_X(x) \asymp 1$) ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  ($P_X(x) \gg n^{-1}$) ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  ($n^{-2} \ll P_X(x) \ll n^{-1}$)
- Hard — w.h.p. the data set has no repeated X's ($P_X(x) \ll n^{-2}$)
  ↘ indistinguishable from nonatomic, so width is $\asymp 1$

Beyond nonatomic

Returning to inference for regression (conf. int. for $\mathbb{E}[Y \mid X]$)...

Suppose P_X is instead *discrete*.
At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  ($P_X(x) \asymp 1$) ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  ($P_X(x) \gg n^{-1}$) ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  ($n^{-2} \ll P_X(x) \ll n^{-1}$) ↘ d.f. inference is still possible!
- Hard — w.h.p. the data set has no repeated X's ($P_X(x) \ll n^{-2}$)
  ↘ indistinguishable from nonatomic, so width is $\asymp 1$

Beyond nonatomic

P might be discrete, nonatomic, or a mixture —
how to unify these three cases?

$M_\gamma(P_X) =$ minimum # of points needed to capture $\geq 1-\gamma$ probability

[18] Lee & B. 2021, *Distribution-free inference for regression: discrete, continuous, and in between*

Beyond nonatomic

P might be discrete, nonatomic, or a mixture —
how to unify these three cases?

$M_\gamma(P_X) =$ minimum # of points needed to capture $\geq 1-\gamma$ probability

Theorem:[18] $\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right] \gtrsim \min\left\{\frac{\left(M_\gamma(P_X)\right)^{1/4}}{n^{1/2}}, 1\right\}$

↗ Vanishing width iff $M_\gamma(P_X) \ll n^2$

[18] Lee & B. 2021, *Distribution-free inference for regression: discrete, continuous, and in between*

Beyond nonatomic

P might be discrete, nonatomic, or a mixture —
how to unify these three cases?

$M_\gamma(P_X) =$ minimum # of points needed to capture $\geq 1 - \gamma$ probability

Theorem:[18] $\mathbb{E}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right] \gtrsim \min\left\{\frac{\left(M_\gamma(P_X)\right)^{1/4}}{n^{1/2}}, 1\right\}$

↗
Vanishing width iff $M_\gamma(P_X) \ll n^2$

(An approx. matching upper bound can be constructed if assume we can accurately estimate the support of P_X and the function $\mathbb{E}[Y \mid X = x]$)

[18] Lee & B. 2021, *Distribution-free inference for regression: discrete, continuous, and in between*

Open questions & future directions

Open questions — algorithms

- Statistically efficient algorithms for the not-nonatomic case, for conditional prediction / marginal inference on μ_P

Open questions & future directions

Open questions — algorithms

- Statistically efficient algorithms for the not-nonatomic case,
  for conditional prediction / marginal inference on μ_P
- Computationally efficient versions of conformal / jackknife+,
  when model alg. is expensive / when Y is multidimensional / etc

Open questions & future directions

Open questions — algorithms

- Statistically efficient algorithms for the not-nonatomic case,
  for conditional prediction / marginal inference on μ_P
- Computationally efficient versions of conformal / jackknife+,
  when model alg. is expensive / when Y is multidimensional / etc
- Can we use the data to guide choices (e.g., score function $S(x, y)$),
  without the need for an additional split of the training data?

Open questions & future directions

Open questions — framework & definitions

- Are there interesting weaker definitions of validity,
  that are still meaningful without assumptions?
  - relaxations of conditional validity, for prediction
  - relaxations of marginal coverage, for inference on regression (e.g., surrogate functions)
  - relaxations of calibration, to allow for continuous predictions

[19] Shah & Peters 2018, *The hardness of conditional independence testing and the generalised covariance measure*

Open questions & future directions

Open questions — framework & definitions

- Are there interesting weaker definitions of validity, that are still meaningful without assumptions?
  - — relaxations of conditional validity, for prediction
  - — relaxations of marginal coverage, for inference on regression (e.g., surrogate functions)
  - — relaxations of calibration, to allow for continuous predictions
- Are there meaningful ways to study distrib.-free hypothesis tests? (Known: impossible to test $X \perp\!\!\!\perp Y \mid Z$ distrib.-free, if Z nonatomic)[19]

[19] Shah & Peters 2018, *The hardness of conditional independence testing and the generalised covariance measure*

Open questions & future directions

Open questions — framework & definitions

- Are there interesting weaker definitions of validity,
  that are still meaningful without assumptions?
  - relaxations of conditional validity, for prediction
  - relaxations of marginal coverage, for inference on regression (e.g., surrogate functions)
  - relaxations of calibration, to allow for continuous predictions
- Are there meaningful ways to study distrib.-free hypothesis tests? (Known: impossible to test $X \perp\!\!\!\perp Y \mid Z$ distrib.-free, if Z nonatomic)[19]
- Are there methods that achieve a weak property for all P, & a stronger property for "nice" P?

[19] Shah & Peters 2018, *The hardness of conditional independence testing and the generalised covariance measure*