Conformal Inference for Frequency Estimation with Sketched Data

Matteo Sesia

Department of Data Sciences and Operations
University of Southern California, Marshall School of Business

Università Cattolica del Sacro Cuore — May 18, 2023

Big data

Large data set of discrete objects Z_i:

$$Z_1, \ldots, Z_n \in \mathscr{Z}$$

For example,

- word tokens in natural language processing [Goyal et al., 2012]
- DNA sequences in genetics [Zhang et al., 2014]
- user features in machine learning, ...

However, storing the entire data set may be unfeasible, due to:

- memory limitations
- privacy concerns

Instead, only a (*sketch*) of the data can be stored.

Frequency estimation from sketched data

Consider the problem of recovering the true frequency of an object $z \in \mathscr{Z}$ within the data set $Z_1, \ldots, Z_n \in \mathscr{Z}$:

$$f_n(z) = \sum_{i=1}^{n} \mathbb{1}\left[Z_i = z\right],$$

using only the information contained in a (lossy) sketch $\mathcal{S}$:

$$\mathcal{S} = \mathcal{S}(Z_1, \ldots, Z_n) \in \mathbb{N}^L,$$

with $L \ll n$.

In general, exact recovery may be impossible. Further, a sensible estimate should depend on how the data are sketched.

The count-min sketch

The count-min sketch (CMS) [Cormode and Muthukrishnan, 2005] is an efficient data structure meant to facilitate the estimation of discrete object frequencies.

The CMS utilizes $d \geq 1$ distinct *hash* functions,

$$h_j : \mathscr{Z} \to [w] = \{1, \ldots, w\},$$

with width $w > 1$, for all $j \in [d] = \{1, \ldots, d\}$.

The data are compressed into a sketch matrix $C \in \mathbb{N}^{d \times w}$:

$$C_{j,k} = \sum_{i=1}^{n} \mathbb{1}\left[h_j(Z_i) = k\right], \qquad j \in [d],\, k \in [w].$$

Therefore, the size of the sketch is $L = d \cdot w \ll n$.

Frequency estimation with the CMS

A classical estimate of $f_n(z)$ for any object $z \in \mathscr{Z}$ is:

$$\hat{f}_{\mathrm{up}}^{\mathrm{CMS}}(z) = \min_{j \in [d]} \left\{ C_{j,h_j(z)} \right\}.$$

This gives a deterministic upper bound for $f_n(z)$ [Cormode and Muthukrishnan, 2005]:

$$\hat{f}_{\mathrm{up}}^{\mathrm{CMS}}(z) \geq f_n(z) = \sum_{i=1}^{n} \mathbb{1}\left[Z_i = z\right].$$

The problem is that this estimate is not necessarily accurate: it is possible that $\hat{f}_{\mathrm{up}}^{\mathrm{CMS}}(z) > f_n(z)$ due random hash collisions.

Probabilistic lower bounds for the CMS

A *pairwise independent* family $\mathcal{H}$ of hash functions is defined as follows. For any $z_1 \neq z_2$ and $h_1, h_2 \sim \mathcal{H}$,

$$\mathbb{P}_{\mathcal{H}}[h_1(z_1) = h_2(z_2)] = \frac{1}{w^2}.$$

Probabilistic lower bounds for the CMS

A *pairwise independent* family $\mathcal{H}$ of hash functions is defined as follows. For any $z_1 \neq z_2$ and $h_1, h_2 \sim \mathcal{H}$,

$$\mathbb{P}_{\mathcal{H}}[h_1(z_1) = h_2(z_2)] = \frac{1}{w^2}.$$

Theorem ([Cormode and Muthukrishnan, 2005])

Suppose the hash functions are chosen at random from a pairwise independent family $\mathcal{H}$.
For any $\delta, \epsilon \in (0,1)$, choosing $d = \lceil -\log \delta \rceil$ and $w = \lceil e/\epsilon \rceil$, for any fixed $z \in \mathscr{Z}$,

$$\mathbb{P}_{\mathcal{H}}[f_n(z) \geq \hat{f}_{\mathrm{up}}^{\mathrm{CMS}}(z) - \epsilon n] \geq 1 - \delta.$$

E.g., if $\delta = 0.05$ and $d = 3$, a 95% lower bound for $f_n(z)$ is:

$$\hat{f}_{\mathrm{up}}^{\mathrm{CMS}}(z) - n \cdot \lceil e/w \rceil.$$

Note: the randomness comes from the hash functions!

Limitations of probabilistic lower bounds for the CMS

- Often too conservative to be useful [Ting, 2018].
- Not very flexible: δ cannot be chosen by the practitioner because it is fixed by d (since $d = \lceil -\log \delta \rceil$), and ϵ is uniquely determined by the hash width (since $w = \lceil e/\epsilon \rceil$).

Limitations of probabilistic lower bounds for the CMS

- Often too conservative to be useful [Ting, 2018].
- Not very flexible: δ cannot be chosen by the practitioner because it is fixed by d (since $d = \lceil -\log \delta \rceil$), and ϵ is uniquely determined by the hash width (since $w = \lceil e/\epsilon \rceil$).

One source of this difficulty is that we have not made any assumptions on the data.

The goal of this work is to develop a more powerful data-driven method to construct flexible "confidence intervals" for $f_n(z)$.

Related work

More recent works have developed uncertainty estimation methods leveraging the randomness in the data.

Bayesian non-parametric approaches:

- [Cai et al., 2018], Dirichlet process on the data distribution.
- Follow-up works: [Dolera et al., 2021], [Favaro and S., 2022]

Frequentist approaches

- [Ting, 2018], resampling (bootstrap) method

Some limitations:

- Model based (or involving some heuristics)
- Specific to the CMS (a linear sketch)

Non-linear sketches

The CMS with *conservative updates* (CMS-CU) mitigates the impact of hash collisions but is not linear.

For each data point Z_i and each $j \in [d]$, update only $C_{j^*(i),k}$:

$$C^{\mathrm{CU}}_{j^*(i),k} \leftarrow C^{\mathrm{CU}}_{j^*(i),k} + \mathbb{1}\left[h_{j^*(i)}(Z_i) = k\right], \quad j^*(i) = \arg\min_{j \in [d]} C_{j,h_j(Z_i)}.$$

Non-linear sketches

The CMS with *conservative updates* (CMS-CU) mitigates the impact of hash collisions but is not linear.

For each data point Z_i and each $j \in [d]$, update only $C_{j^*(i),k}$:

$$C^{\mathrm{CU}}_{j^*(i),k} \leftarrow C^{\mathrm{CU}}_{j^*(i),k} + \mathbb{1}\left[h_{j^*(i)}(Z_i) = k\right], \quad j^*(i) = \arg\min_{j \in [d]} C_{j,h_j(Z_i)}.$$

Then, return as usual:

$$\hat{f}^{\mathrm{CMS-CU}}_{\mathrm{up}}(z) = \min_{j \in [d]} \left\{ C^{\mathrm{CU}}_{j,h_j(z)} \right\}.$$

Non-linear sketches

The CMS with *conservative updates* (CMS-CU) mitigates the impact of hash collisions but is not linear.

For each data point Z_i and each $j \in [d]$, update only $C_{j^*(i),k}$:

$$C^{\mathrm{CU}}_{j^*(i),k} \leftarrow C^{\mathrm{CU}}_{j^*(i),k} + \mathbb{1}\left[h_{j^*(i)}(Z_i) = k\right], \quad j^*(i) = \arg\min_{j \in [d]} C_{j,h_j(Z_i)}.$$

Then, return as usual:

$$\hat{f}^{\mathrm{CMS-CU}}_{\mathrm{up}}(z) = \min_{j \in [d]} \left\{ C^{\mathrm{CU}}_{j,h_j(z)} \right\}.$$

This guarantees:

- $\hat{f}^{\mathrm{CMS-CU}}_{\mathrm{up}}(z) \geq f_n(z)$
- $\hat{f}^{\mathrm{CMS-CU}}_{\mathrm{up}}(z) \leq \hat{f}^{\mathrm{CMS}}_{\mathrm{up}}(z)$

Desiderata

We would like to construct "confidence intervals" for $f_n(z)$ that:

1. do not require knowing the distribution of the data
2. are not limited to the specific linear structure of the CMS
3. are provably valid in finite samples
4. avoid being too conservative by adapting to the observed data

Desiderata

We would like to construct "confidence intervals" for $f_n(z)$ that:

1. do not require knowing the distribution of the data
2. are not limited to the specific linear structure of the CMS
3. are provably valid in finite samples
4. avoid being too conservative by adapting to the observed data

Key assumption:

$$Z_1, Z_2, \ldots, Z_n \sim P_Z$$

Outline

1. Review of conformal prediction
2. Method: conformalized sketching
3. Numerical experiments
4. Discussion

Exchangeable random variables

We say that $Z_1, Z_2, \ldots, Z_n$ are exchangeable if and only if, for any permutation σ of $\{1, \ldots, n\}$,

$$p(Z_1, Z_2, \ldots, Z_n) = p(Z_{\sigma(1)}, Z_{\sigma(2)}, \ldots, Z_{\sigma(n)}).$$

For example, $Z_1, Z_2, \ldots, Z_n$ are exchangeable if they are i.i.d.

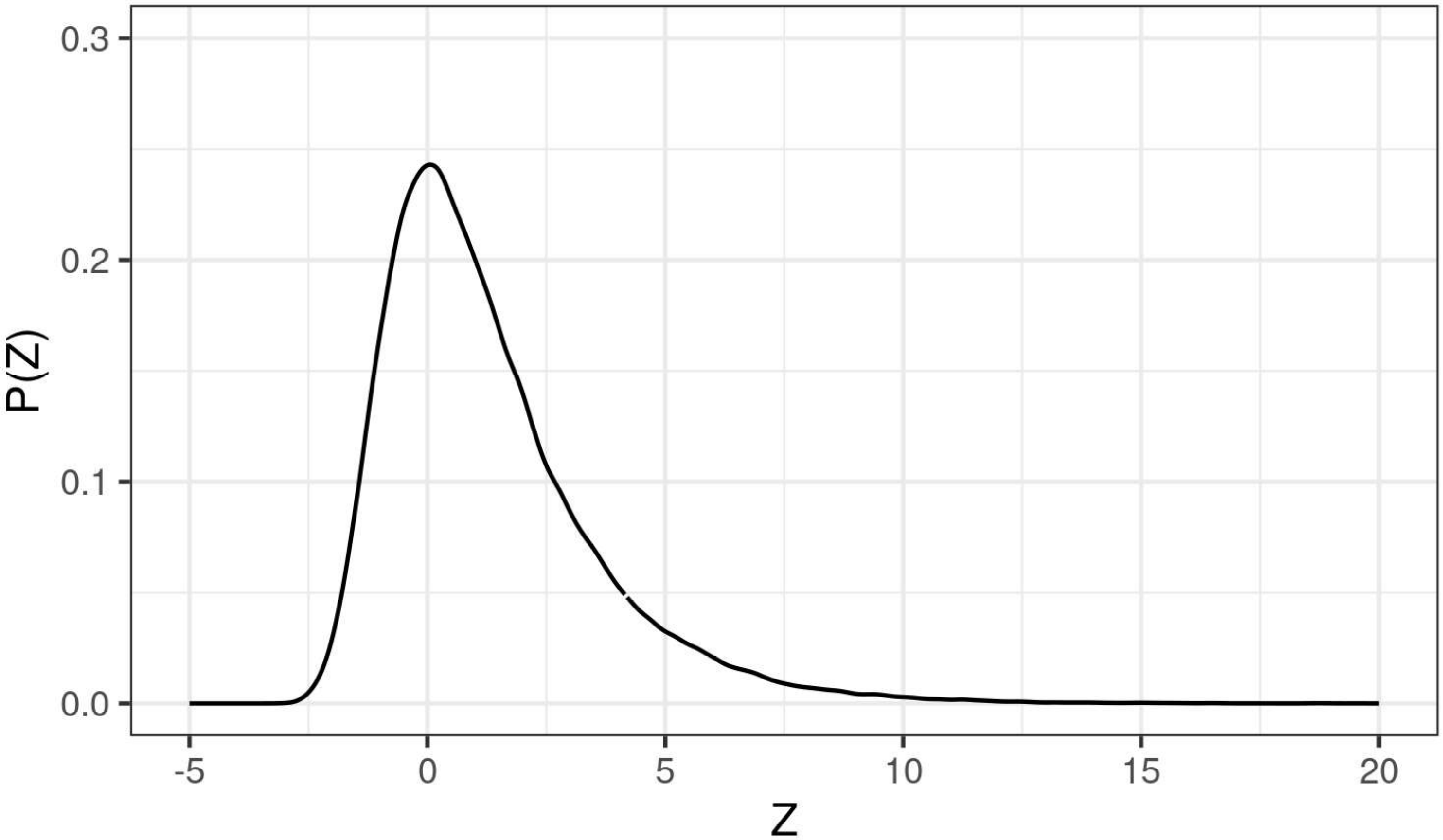
Conformal prediction without covariates

Suppose we have

$$Z_i \overset{\text{exch.}}{\sim} P_Z, \qquad Z \in \mathbb{R}$$

and we want to use the first n data points to construct a one-sided prediction interval $\hat{C}_\alpha = (-\infty, \hat{U}_{1-\alpha}]$ such that

$$\mathbb{P}\left[Z_{n+1} \leq \hat{U}_{1-\alpha}\right] \geq 1 - \alpha.$$



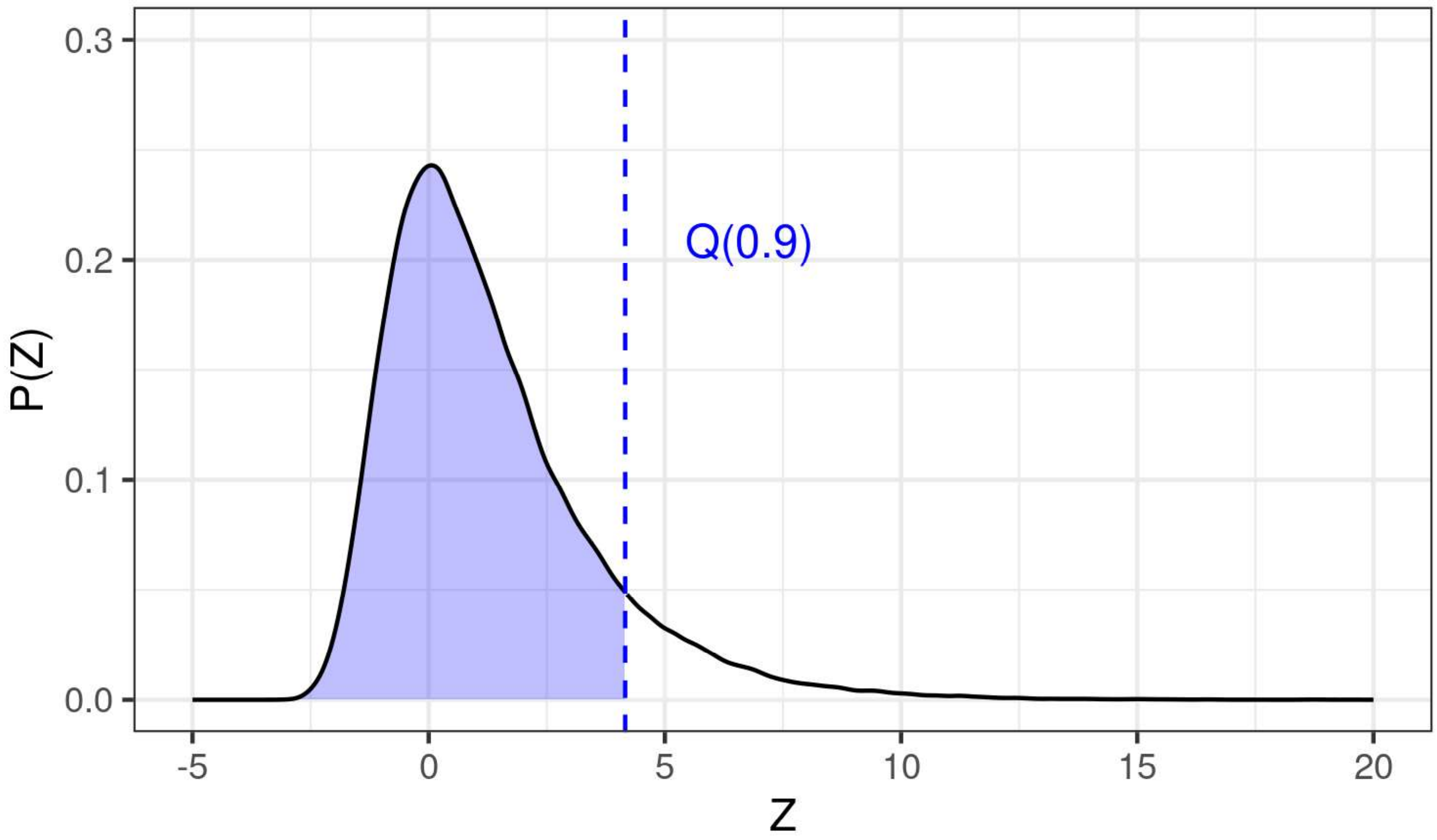
Conformal prediction without covariates

Suppose we have

$$Z_i \overset{\text{exch.}}{\sim} P_Z, \qquad Z \in \mathbb{R}$$

and we want to use the first n data points to construct a one-sided prediction interval $\hat{C}_\alpha = (-\infty, \hat{U}_{1-\alpha}]$ such that

$$\mathbb{P}\left[Z_{n+1} \leq \hat{U}_{1-\alpha}\right] \geq 1 - \alpha.$$



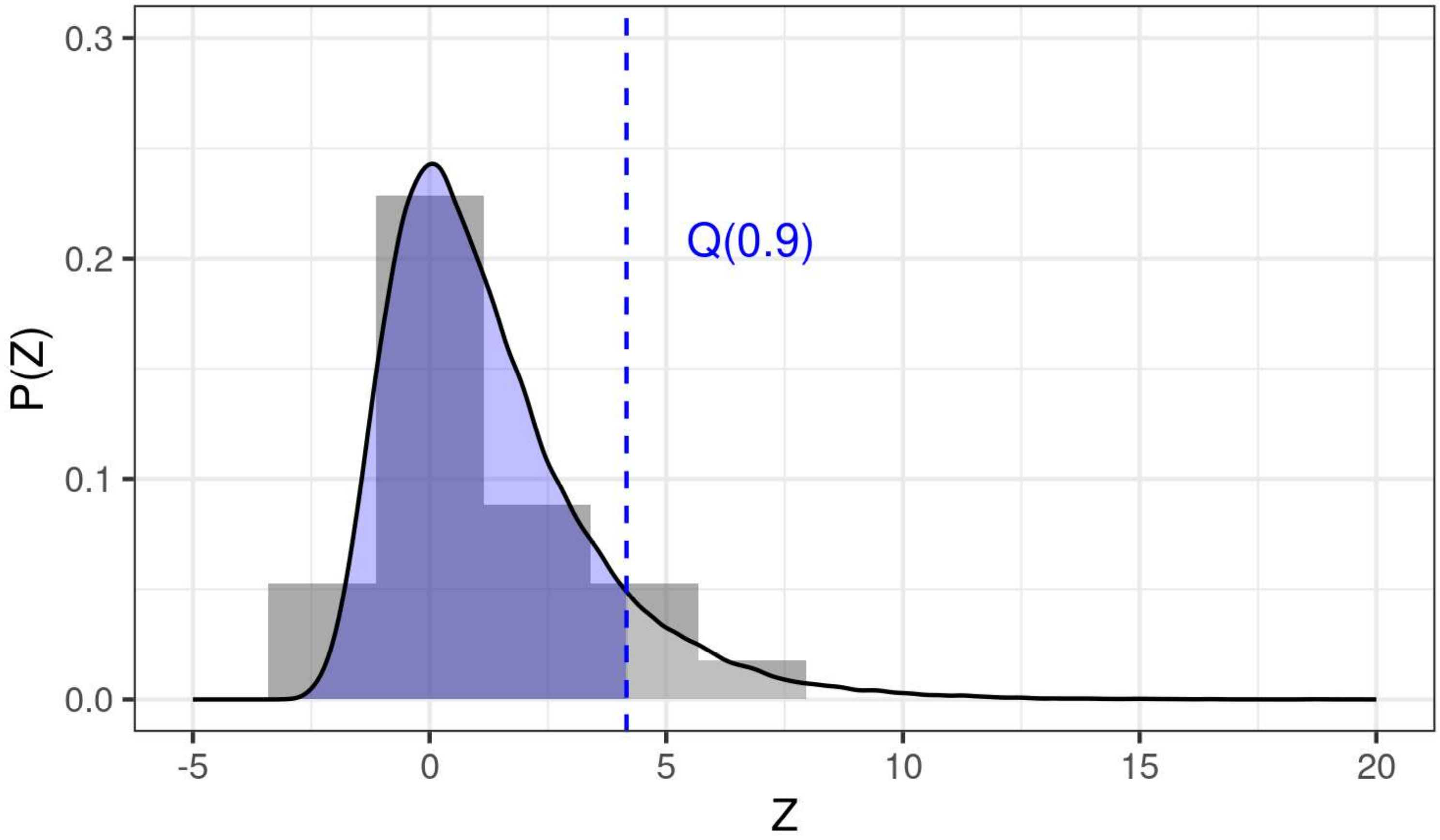
Conformal prediction without covariates

Suppose we have

$$Z_i \overset{\text{exch.}}{\sim} P_Z, \qquad Z \in \mathbb{R}$$

and we want to use the first n data points to construct a one-sided prediction interval $\hat{C}_\alpha = (-\infty, \hat{U}_{1-\alpha}]$ such that

$$\mathbb{P}\left[Z_{n+1} \leq \hat{U}_{1-\alpha}\right] \geq 1 - \alpha.$$



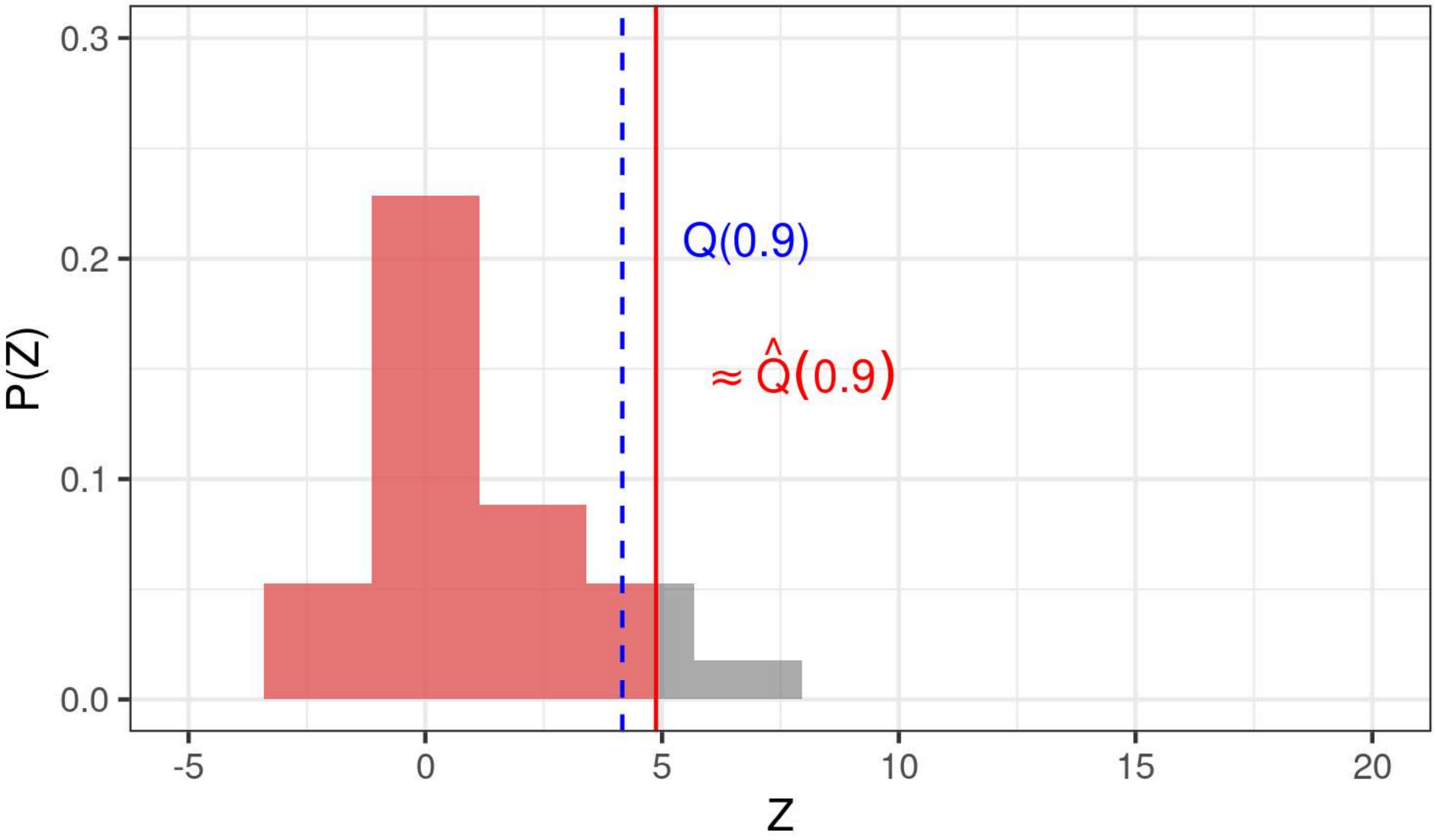
Conformal prediction without covariates

Suppose we have

$$Z_i \overset{\text{exch.}}{\sim} P_Z, \qquad Z \in \mathbb{R}$$

and we want to use the first n data points to construct a one-sided prediction interval $\hat{C}_\alpha = (-\infty, \hat{U}_{1-\alpha}]$ such that

$$\mathbb{P}\left[Z_{n+1} \leq \hat{U}_{1-\alpha}\right] \geq 1 - \alpha.$$



Finite-sample inflation of empirical quantiles

Empirical CDF and quantile function:

$$\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[Z_i \leq z\right], \qquad \hat{Q}_n(\alpha) = Z_{(\lceil \alpha n \rceil)}$$

Lemma (E.g., [Vovk et al., 2005, Romano et al., 2019])

Suppose $Z_1, \ldots, Z_{n+1}$ are exchangeable random variables. For any $\alpha \in \{0, 1\}$, define $\alpha_n = \left(1 + \frac{1}{n}\right)\alpha$. Then,

$$\mathbb{P}\left[Z_{n+1} \leq \hat{Q}_n(\alpha_n)\right] \leq \alpha.$$

Moreover, if $\{Z_1, \ldots, Z_{n+1}\}$ are a.s. distinct,

$$\mathbb{P}\left[Z_n \leq \hat{Q}_n(\alpha_n)\right] \leq \alpha + \frac{1}{n+1}.$$

One-sided conformal prediction without covariates

Suppose $Z_1, \ldots, Z_{n+1}$ are exchangeable random variables.
For any $\alpha \in \{0, 1\}$, define $\hat{C}_\alpha$ as

$$\hat{C}_\alpha = (-\infty, \hat{Q}_n(\alpha_n)].$$

Then,

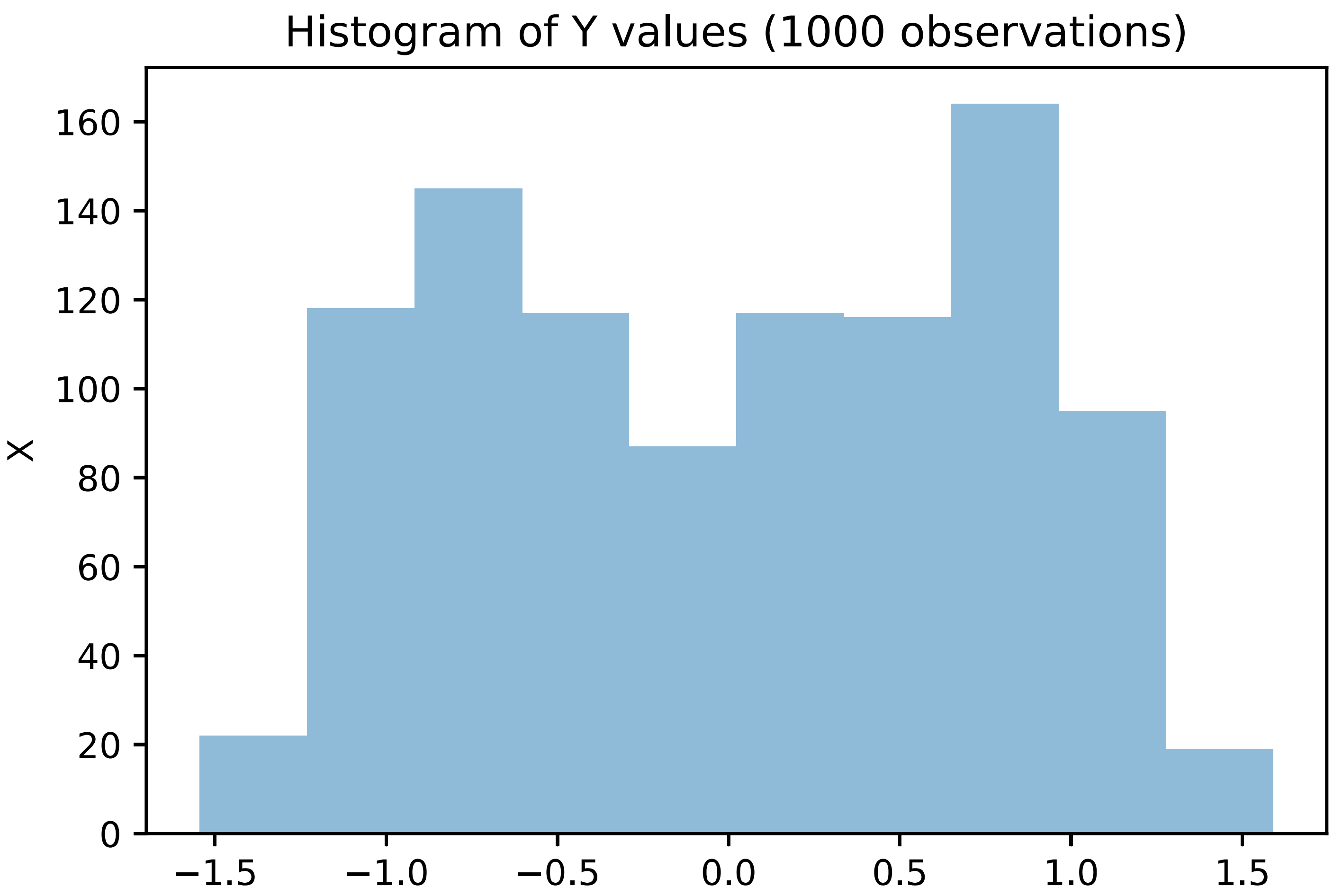
$$\alpha \leq \mathbb{P}\left[Z_{n+1} \in \hat{C}_\alpha\right] \leq \alpha + \frac{1}{n}.$$

Conformal prediction with covariates

Suppose we have

$$(X_i, Y_i)_{i=1}^{n+1} \overset{\text{exch.}}{\sim} P_{X,Y}, \qquad X \in \mathbb{R}^p, Y \in \mathbb{R}$$

We would like to predict Y_{n+1} given $(X_i, Y_i)_{i=1}^{n}$

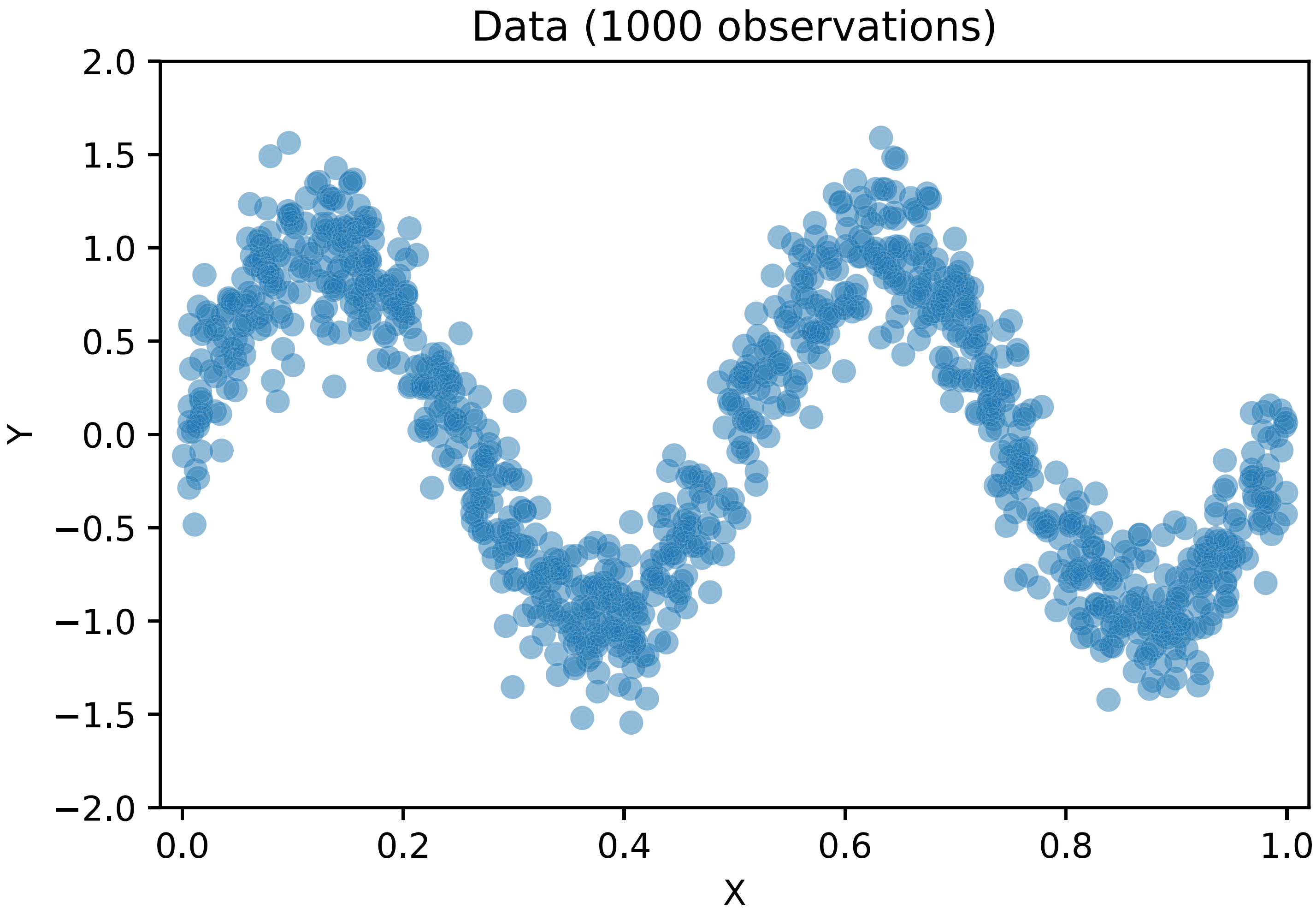


Conformal prediction with covariates

Suppose we have

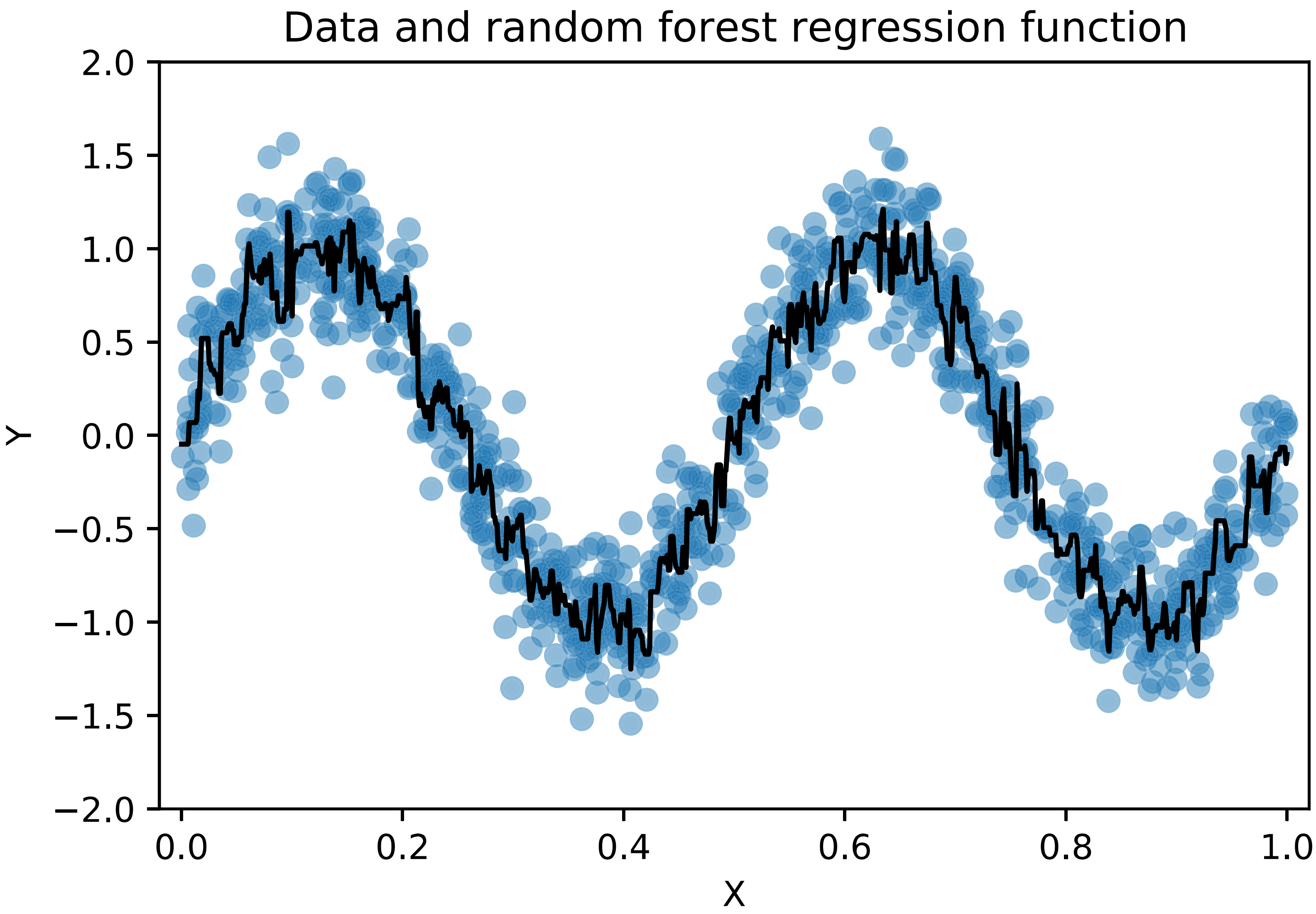
$$(X_i, Y_i)_{i=1}^{n+1} \overset{\text{exch.}}{\sim} P_{X,Y}, \qquad X \in \mathbb{R}^p, Y \in \mathbb{R}$$

We would like to predict Y_{n+1} given $(X_i, Y_i)_{i=1}^{n}$ and X_{n+1}



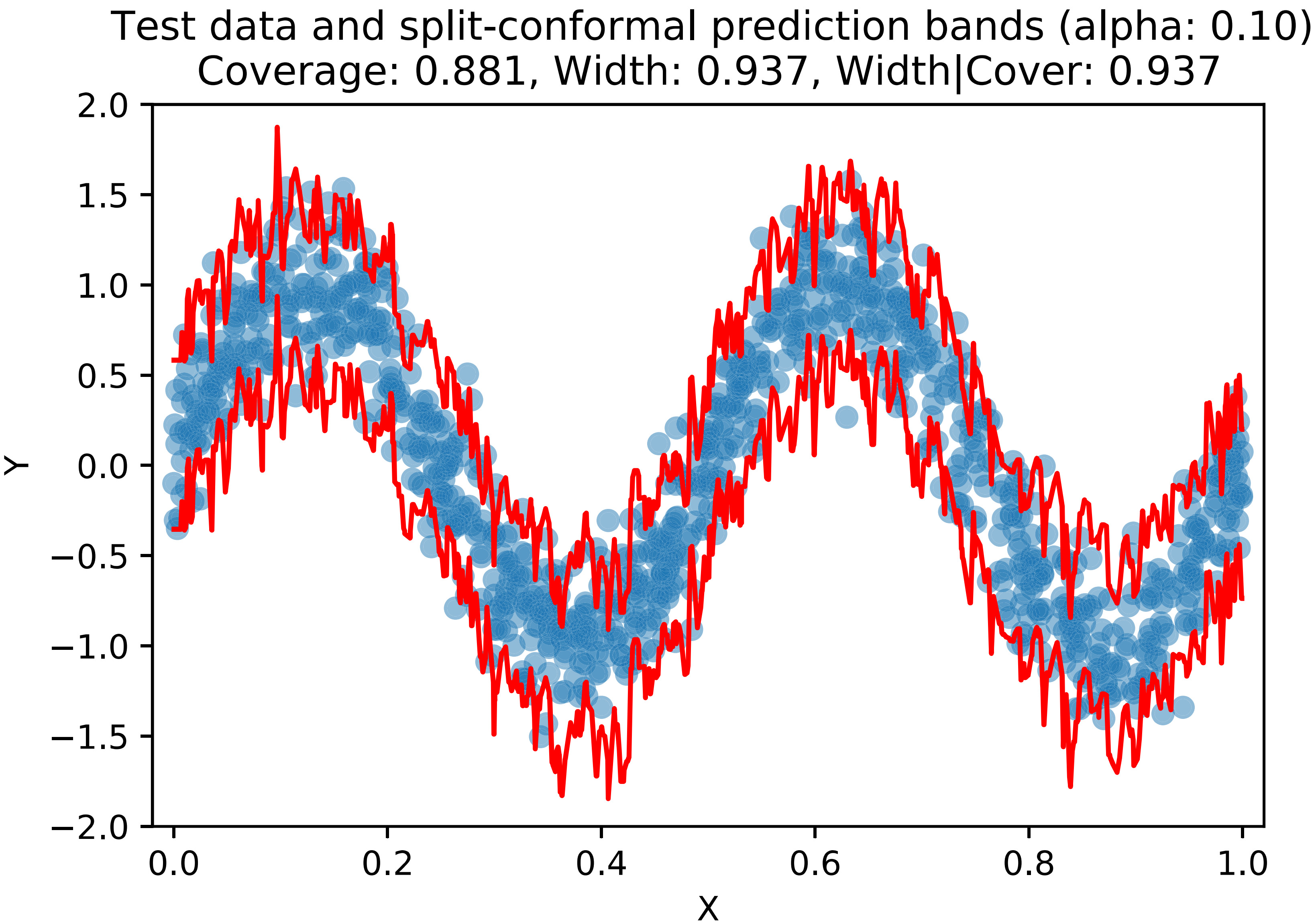
Machine-learning prediction

Lots of machine-learning algorithms. But how confident are we?



Machine-learning prediction

Lots of machine-learning algorithms. But how confident are we?



Conformal prediction

Key ideas:

1. Use ML to project project the problem into 1 dimension.
2. Apply the empirical quantile lemmas presented earlier.
3. Some kind of data hold-out is needed to ensure exchangeability with the test data.

This is a general recipe, many different variations are possible.

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, $\alpha \in (0, 1)$
black-box model $\mathcal{B}$, level $\alpha \in (0, 1)$

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^{n}$, test point X_{n+1}, $\alpha \in (0,1)$
black-box model $\mathcal{B}$, level $\alpha \in (0,1)$

Split the data: $\mathcal{I}_1 = \{1, \ldots, n/2\}, \mathcal{I}_2 = \{n/2+1, \ldots, n\}$

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, $\alpha \in (0,1)$
black-box model $\mathcal{B}$, level $\alpha \in (0,1)$

Split the data: $\mathcal{I}_1 = \{1, \ldots, n/2\}, \mathcal{I}_2 = \{n/2+1, \ldots, n\}$
Train $\mathcal{B}$ on $\mathcal{I}_1 : \mathcal{B}\left(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}\right) \to \hat{f}$

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, $\alpha \in (0,1)$
black-box model $\mathcal{B}$, level $\alpha \in (0,1)$

Split the data: $\mathcal{I}_1 = \{1, \ldots, n/2\}, \mathcal{I}_2 = \{n/2+1, \ldots, n\}$
Train $\mathcal{B}$ on $\mathcal{I}_1$: $\mathcal{B}\left(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}\right) \to \hat{f}$
Evaluate residuals on $\mathcal{I}_2$: $Z_i = \left|Y_i - \hat{f}(X_i)\right|$, for all $i \in \mathcal{I}_2$

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, $\alpha \in (0,1)$
black-box model $\mathcal{B}$, level $\alpha \in (0,1)$

Split the data: $\mathcal{I}_1 = \{1, \ldots, n/2\}, \mathcal{I}_2 = \{n/2+1, \ldots, n\}$
Train $\mathcal{B}$ on $\mathcal{I}_1 : \mathcal{B}\left(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}\right) \to \hat{f}$
Evaluate residuals on $\mathcal{I}_2 : Z_i = \left|Y_i - \hat{f}(X_i)\right|$, for all $i \in \mathcal{I}_2$
Compute $\hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n) = Z_{(\lceil n\beta_n \rceil)}$, where $\beta_n = (1-\alpha)(1+1/n)$

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, $\alpha \in (0,1)$

black-box model $\mathcal{B}$, level $\alpha \in (0,1)$

Split the data: $\mathcal{I}_1 = \{1, \ldots, n/2\}, \mathcal{I}_2 = \{n/2+1, \ldots, n\}$

Train $\mathcal{B}$ on $\mathcal{I}_1$: $\mathcal{B}\left(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}\right) \to \hat{f}$

Evaluate residuals on $\mathcal{I}_2$: $Z_i = \left|Y_i - \hat{f}(X_i)\right|$, for all $i \in \mathcal{I}_2$

Compute $\hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n) = Z_{(\lceil n\beta_n \rceil)}$, where $\beta_n = (1-\alpha)(1+1/n)$

Output:

$\hat{C}_\alpha(X_{n+1}) = [\hat{f}(X_{n+1}) - \hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n), \hat{f}(X_{n+1}) + \hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n)]$

Split-conformal prediction [Vovk et al., 2005]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, $\alpha \in (0,1)$
black-box model $\mathcal{B}$, level $\alpha \in (0,1)$

Split the data: $\mathcal{I}_1 = \{1, \ldots, n/2\}, \mathcal{I}_2 = \{n/2+1, \ldots, n\}$
Train $\mathcal{B}$ on $\mathcal{I}_1$: $\mathcal{B}\left(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}\right) \to \hat{f}$
Evaluate residuals on $\mathcal{I}_2$: $Z_i = \left|Y_i - \hat{f}(X_i)\right|$, for all $i \in \mathcal{I}_2$
Compute $\hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n) = Z_{(\lceil n\beta_n \rceil)}$, where $\beta_n = (1-\alpha)(1+1/n)$

Output:
$\hat{C}_\alpha(X_{n+1}) = [\hat{f}(X_{n+1}) - \hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n), \hat{f}(X_{n+1}) + \hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n)]$

Why does this work?

$$Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \quad \Longleftrightarrow \quad Z_{n+1} \leq \hat{Q}_n(Z_{\mathcal{I}_2}, \beta_n).$$

Outline

1. Review of conformal prediction
2. Method: conformalized sketching
3. Numerical experiments
4. Discussion

Problem setup

Key assumption:

$$Z_1, \ldots, Z_n, Z_{n+1} \overset{\text{exch.}}{\sim} P_Z$$

Sketch $Z_1, \ldots, Z_n \to \phi_n = \phi(Z_1, \ldots, Z_n)$.

Then, estimate $f_n(Z_{n+1})$ using ϕ_n, where

$$f_n(z) = \sum_{i=1}^{n} \mathbb{1}\left[Z_i = z\right], \qquad \forall z \in \mathscr{Z}.$$

Problem setup

Key assumption:

$$Z_1, \ldots, Z_n, {\color{red}Z_{n+1}} \overset{\text{exch.}}{\sim} P_Z$$

Sketch $Z_1, \ldots, Z_n \to \phi_n = \phi(Z_1, \ldots, Z_n)$.

Then, estimate $f_n({\color{red}Z_{n+1}})$ using ϕ_n, where

$$f_n(z) = \sum_{i=1}^{n} \mathbb{1}\left[Z_i = z\right], \qquad \forall z \in \mathscr{Z}.$$

Construct a (tight) prediction interval

$$[\hat{L}_{n,\alpha}(Z_{n+1}; \phi_n), \hat{U}_{n,\alpha}(Z_{n+1}; \phi_n)]$$

with guaranteed *marginal coverage*:

$$\mathbb{P}\left[\hat{L}_{n,\alpha}(Z_{n+1}; \phi_n) \leq f_n(Z_{n+1}) \leq \hat{U}_{n,\alpha}(Z_{n+1}; \phi_n)\right] \geq 1 - \alpha.$$

for any fixed $\alpha \in (0, 1)$,

Step 1: warm-up

During an initial *warm-up* phase, the frequencies of the n_0 distinct objects among the first $m \ll n$ observations from the data stream, $Z_1, \ldots, Z_{n_0}$, are stored exactly into f_m,

$$f_m^{\mathrm{wu}}(z) = \sum_{i=0}^{m} \mathbb{1}\left[Z_i = z\right]. \tag{1}$$

Storage requirement: $\mathcal{O}(n_0) \leq \mathcal{O}(m) \ll \mathcal{O}(n)$.

Step 2: sketching

The remaining $m - m$ data points are streamed and sketched, storing also the true frequencies for all instances of objects already seen during the warm-up phase.

Sketch:

$$\phi(Z_{m+1}, \dots, Z_n)$$

.

The following counters are also computed and stored:

$$f^{\mathrm{sv}}_{n-m}(z) = \begin{cases} \sum_{i=m+1}^{n} \mathbb{1}\left[Z_i = z\right], & \text{if } f^{\mathrm{wu}}_m(z) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Again, the memory cost is $\ll O(n)$.

Step 3: conformalization

For all $i \in \{1, \ldots, m\} \cup \{n+1\}$, define

$$Y_i = \sum_{i'=m+1}^{n} \mathbb{1}\left[Z_{i'} = Z_i\right],$$

the true frequency of Z_i among $Z_{m+1}, \ldots, Z_n$.

Note that Y_i is observable for $i \in \{1, \ldots, m\}$, in which case

$$Y_i = f^{\mathrm{sv}}_{n-m}(Z_i).$$

For a new query Z_{n+1}, the target of inference is

$$f_n(Z_{n+1}) = Y_{n+1} + f^{\mathrm{wu}}_m(Z_{n+1}),$$

but the second term is known. So, we just need to predict Y_{n+1}.

Step 3: conformalization (continued)

Next, we need to define meaningful features X.

For each $i \in \{1, \ldots, m\} \cup \{n+1\}$, define

$$X_i = (Z_i, \phi(Z_{m+1}, \ldots, Z_n)) .$$

Step 3: conformalization (continued)

Next, we need to define meaningful features X.

For each $i \in \{1, \ldots, m\} \cup \{n+1\}$, define

$$X_i = (Z_i, \phi(Z_{m+1}, \ldots, Z_n)) .$$

Proposition (S. and Favaro, 2022)

If the data $Z_1, \ldots, Z_{n+1}$ are exchangeable, the pairs of random variables $(X_1, Y_1), \ldots, (X_m, Y_m), (X_{n+1}, Y_{n+1})$ are exchangeable.

Therefore, we can apply conformal prediction to estimate Y_{n+1}.

Conformity scores

Take a *nested* sequence of intervals indexed by $t \in \mathcal{T} \subseteq \mathbb{R}$,

$$[\hat{L}_{m,\alpha}(x; t), \hat{U}_{m,\alpha}(x; t)].$$

Suppose $\exists t_\infty \in \mathcal{T}$ s.t. $\hat{L}_{m,\alpha}(x; t_\infty) \leq Y \leq \hat{U}_{m,\alpha}(x; t_\infty)$ a.s. $\forall x$.

Conformity scores

Take a *nested* sequence of intervals indexed by $t \in \mathcal{T} \subseteq \mathbb{R}$,

$$[\hat{L}_{m,\alpha}(x;t), \hat{U}_{m,\alpha}(x;t)].$$

Suppose $\exists t_\infty \in \mathcal{T}$ s.t. $\hat{L}_{m,\alpha}(x;t_\infty) \le Y \le \hat{U}_{m,\alpha}(x;t_\infty)$ a.s. $\forall x$.

For each $i \in \{1,\ldots,m\}$, compute $E(X_i, Y_i)$, where

$$E(x,y) = \inf\left\{t \in \mathcal{T} : Y \in [\hat{L}_{m,\alpha}(x;t), \hat{U}_{m,\alpha}(x;t)]\right\}.$$

Conformity scores

Take a *nested* sequence of intervals indexed by $t \in \mathcal{T} \subseteq \mathbb{R}$,

$$[\hat{L}_{m,\alpha}(x; t), \hat{U}_{m,\alpha}(x; t)].$$

Suppose $\exists t_\infty \in \mathcal{T}$ s.t. $\hat{L}_{m,\alpha}(x; t_\infty) \leq Y \leq \hat{U}_{m,\alpha}(x; t_\infty)$ a.s. $\forall x$.

For each $i \in \{1, \ldots, m\}$, compute $E(X_i, Y_i)$, where

$$E(x, y) = \inf \left\{ t \in \mathcal{T} : Y \in [\hat{L}_{m,\alpha}(x; t), \hat{U}_{m,\alpha}(x; t)] \right\}.$$

The conformal prediction interval given X_{n+1} is then:

$$\hat{C}_{1-\alpha}(X_{n+1}) = \left[\hat{L}_{m,\alpha}(X_{n+1}; \hat{Q}_{m,1-\alpha}), \hat{U}_{m,\alpha}(X_{n+1}; \hat{Q}_{m,1-\alpha})\right],$$

where $\hat{Q}_{m,1-\alpha}$ is the (inflated) empirical quantile of $E(X_i, Y_i)$.

Note that $Y_{n+1} \notin \hat{C}_{1-\alpha}(X_{n+1}) \iff E(X_{n+1}, Y_{n+1}) > \hat{Q}_{m,1-\alpha}$.

Fixed one-sided conformity scores

In our sketching problem with CMS (or CMS-CU), we already have a deterministic upper bound, $\hat{f}_{n-m,\mathrm{up}}(Z_{n+1})$.

To construct a monotone sequence of lower bounds $\hat{L}_{m,\alpha}(\cdot; t)$, a simple option is to shift the upper bound by a constant:

$$\hat{L}^{\mathrm{fixed}}_{m,\alpha}((z,\phi); t) = \max\{0, \hat{f}_{n-m,\mathrm{up}}(Z_{n+1}) - t\}.$$

This gives the following conformity scores:

$$\begin{aligned} E_i &= \inf\left\{t \in \mathcal{T} : Y_i \in [\hat{f}_{n-m,\mathrm{up}}(Z_i) - t, \hat{f}_{n-m,\mathrm{up}}(Z_i)]\right\} \\ &= \hat{f}_{n-m,\mathrm{up}}(Z_i) - Y_i. \end{aligned}$$

Adaptive one-sided conformity scores

Fit an ML model to estimate the conditional distribution of

$$\hat{f}_{n-m,\mathrm{up}}(Z_i) - Y_i \mid \hat{f}_{n-m,\mathrm{up}}(Z_i),$$

using a subset of $m^{\mathrm{train}} < m$ supervised data points (X_i, Y_i).

E.g., multiple quantile neural network [Taylor, 2000] or a quantile random forest [Meinshausen, 2006].

Let $\hat{q}_t$ be the estimated α_t-th lower conditional quantile, for all $t \in \{1, \dots, T\}$ and some fixed sequence $0 = \alpha_1 < \dots < \alpha_T = 1$.

Without loss of generality, let $\hat{q}_0 = 0$ and $\hat{q}_T = m$.

Then, for $t \in \{0, 1, \dots, m\}$, define

$$\hat{L}^{\mathrm{adaptive}}_{m,\alpha}((z, \phi); t) = \max\left\{0, \hat{f}_{n-m,\mathrm{up}}(X_{n+1}) - \hat{q}_t\left(\hat{f}_{n-m,\mathrm{up}}(X_{n+1})\right)\right\}.$$

This approach can lead to a lower bound whose distance from the upper bound depends on X_{n+1}.

Outline

1. Review of conformal prediction
2. Method: conformalized sketching
3. Numerical experiments
4. Discussion

Performance with Zipf data

Observations: $n = 100,000$ i.i.d. from Zipf(a), with $a > 1$.

$$\mathbb{P}\left[Z_i = z\right] = \frac{z^{-a}}{\zeta(a)}, \text{ for all } z \in \{1, 2 \ldots, \}.$$

Sketch: CMS-CU.

Hash functions: $d = 3$ with width $w = 1000$.

Warm-up observations: $m = 5,000$.

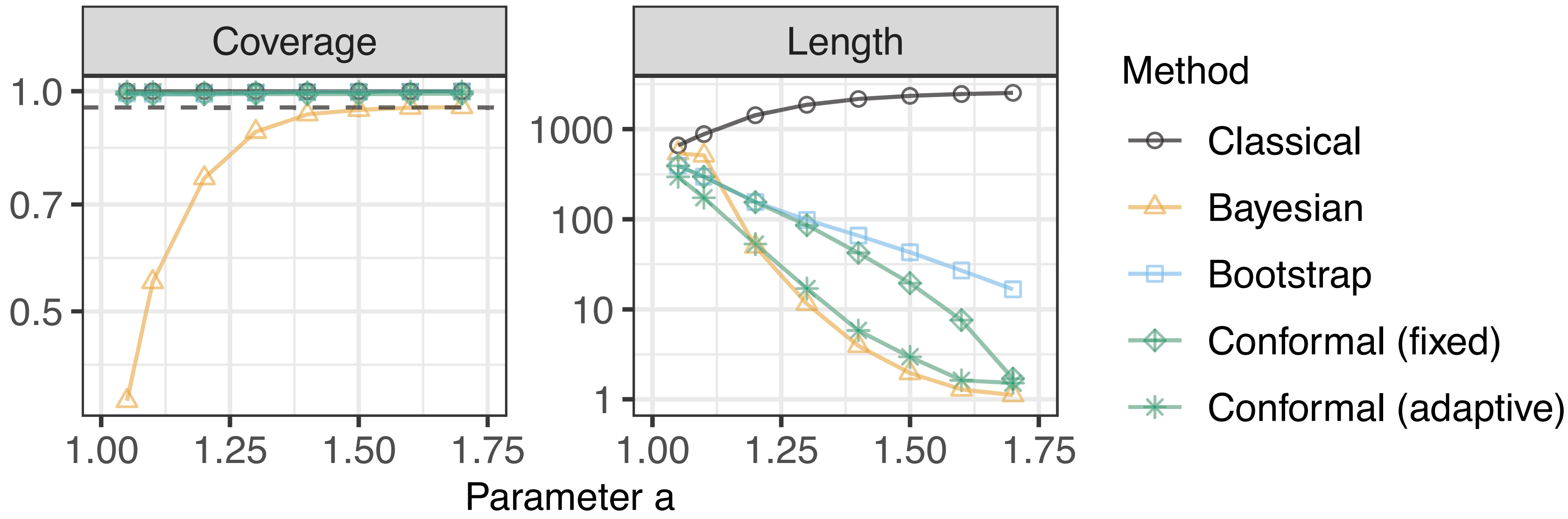
Performance with Zipf data

Observations: $n = 100,000$ i.i.d. from Zipf(a), with $a > 1$.

$$\mathbb{P}\left[Z_i = z\right] = \frac{z^{-a}}{\zeta(a)}, \text{ for all } z \in \{1, 2 \ldots, \}.$$

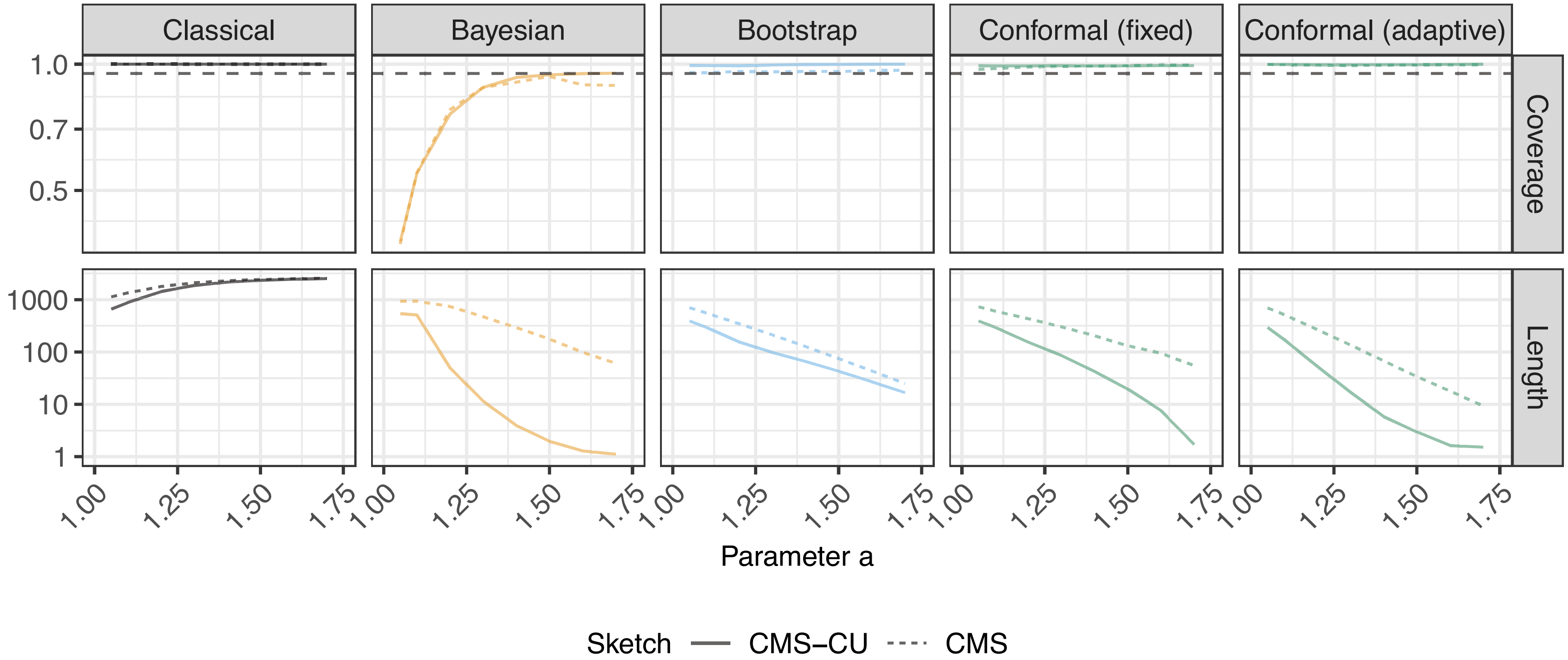
Sketch: CMS-CU.
Hash functions: $d = 3$ with width $w = 1000$.

Warm-up observations: $m = 5,000$.



Performance of 95% confidence intervals with simulated Zipf data sketched with CMS-CU. The results are shown as a function of the Zipf tail parameter a.

Comparison of different sketches

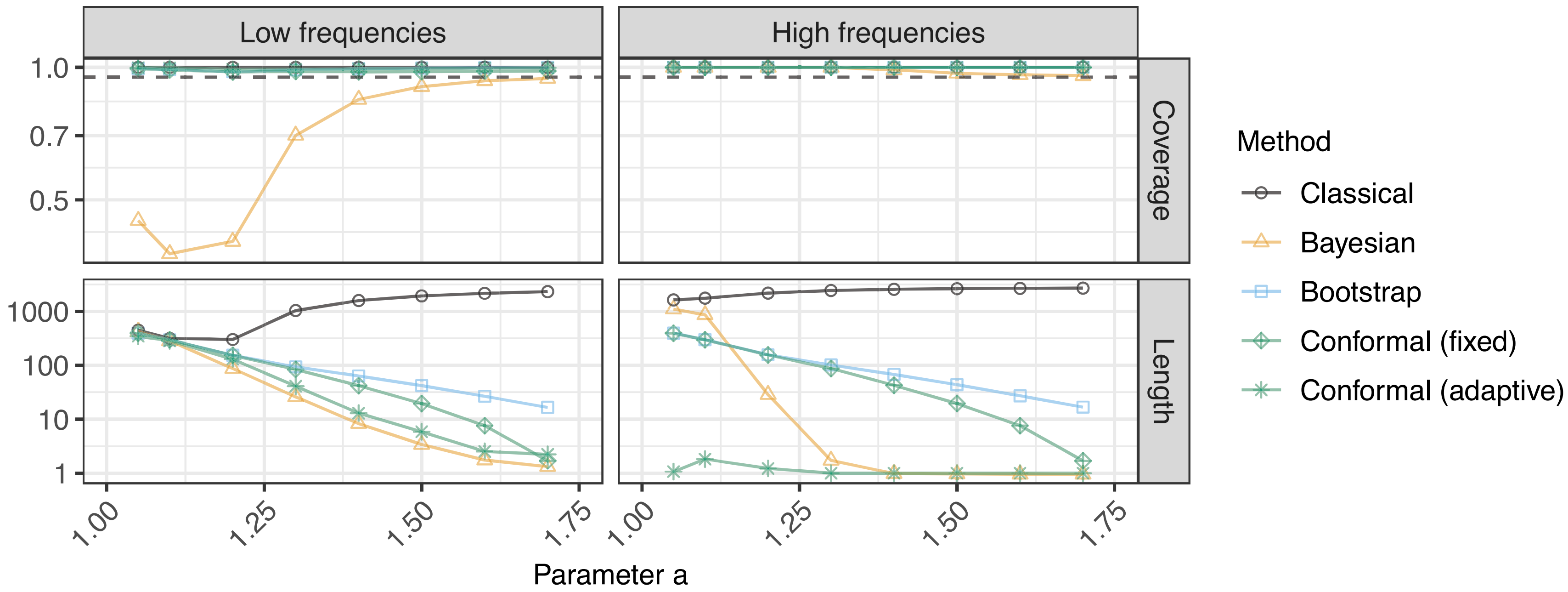


Performance of 95% confidence intervals for random queries, based on synthetic data from a Zipf distribution. The data are sketched with either the vanilla CMS or the CMS-CU. The results are shown as a function of the Zipf tail parameter *a*.

Performance for queries with different frequency

Not all queries are the same.

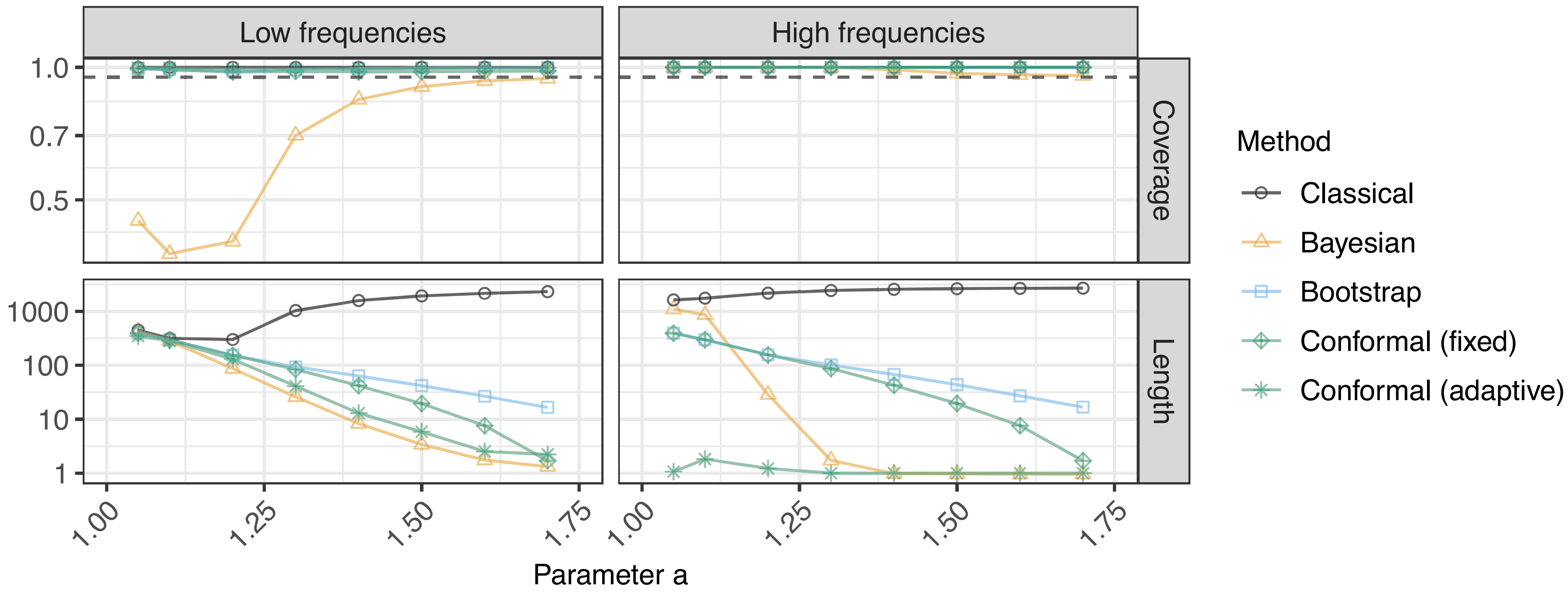
Queries of rarer objects are more difficult.



Performance of 95% confidence intervals stratified by the true query frequency. Left: frequency below median; right: above median.

Performance for queries with different frequency

Not all queries are the same.
Queries of rarer objects are more difficult.



Performance of 95% confidence intervals stratified by the true query frequency. Left: frequency below median; right: above median.

It is possible to guarantee *frequency-range conditional coverage*:

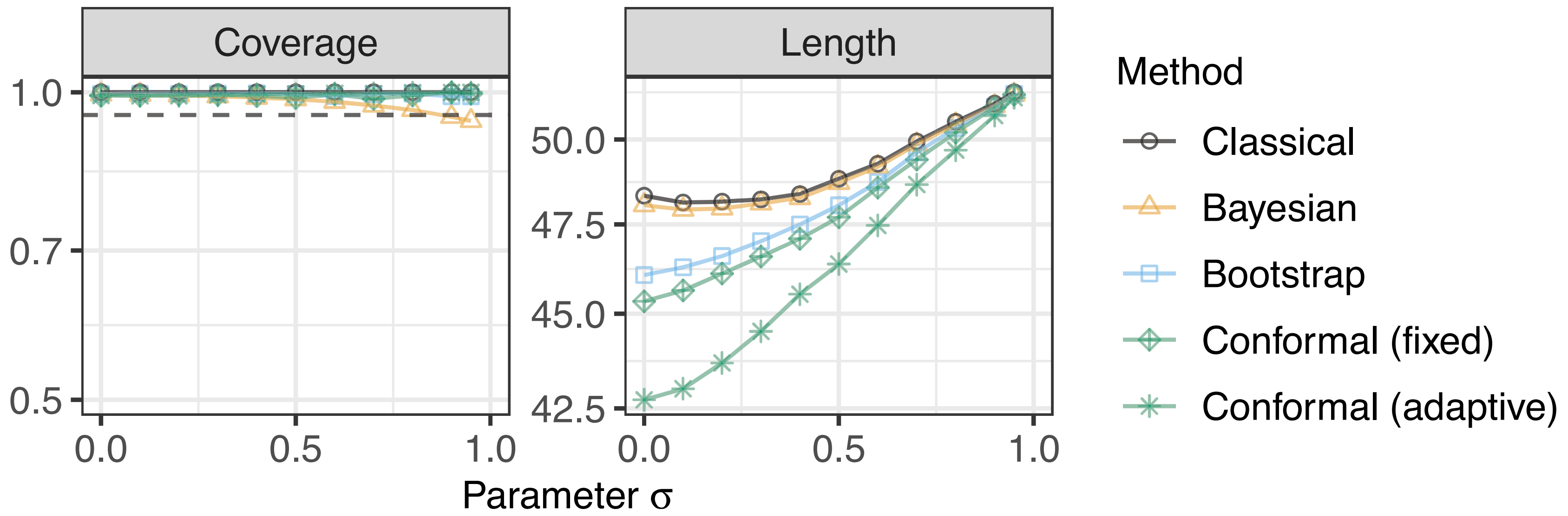
$$\mathbb{P}\left[f_n(Z_{n+1}) \in \hat{C}_{n,\alpha}(Z_{n+1}) \mid f_n(Z_{n+1}) \in B\right] \geq 1 - \alpha, \ \forall B \in \mathcal{B}.$$

Performance with Pitman-Yor Process data

Observations: $n = 100,000$ i.i.d. from $\mathrm{PYP}(5000, \sigma)$ [Pitman and Yor, 1997], with $\sigma \in [0, 1)$. Note that $\sigma = 0$ corresponds to a Dirichlet process, matching the assumption of the Bayesian benchmark.

Performance with Pitman-Yor Process data

Observations: $n = 100,000$ i.i.d. from PYP$(5000, \sigma)$ [Pitman and Yor, 1997], with $\sigma \in [0, 1)$. Note that $\sigma = 0$ corresponds to a Dirichlet process, matching the assumption of the Bayesian benchmark.



Empirical coverage and length of 95% confidence intervals for random queries on synthetic data from the predictive distribution of a Pitman-Yor process. The data are sketched with the CMS-CU. The results are shown as a function of the Pitman-Yor process parameter σ.

Analysis of 2-grams in English literature

Data: 18 open-domain pieces of classic English literature downloaded from the Gutenberg Corpus [Project Gutenberg, sent].

The goal is to count the frequencies of all *2-grams*—consecutive pairs of English words.

After some pre-proccessing, the number of 2-grams is $\approx$ 1,700,000. The total number of all *possible* 2-grams is $\approx$ 650,000,000.

Analysis of 2-grams in English literature

Data: 18 open-domain pieces of classic English literature downloaded from the Gutenberg Corpus [Project Gutenberg, sent].

The goal is to count the frequencies of all *2-grams*—consecutive pairs of English words.

After some pre-proccessing, the number of 2-grams is $\approx$ 1,700,000.
The total number of all *possible* 2-grams is $\approx$ 650,000,000.

Sketch 1,000,000 2-grams, query 10,000 2-grams.
The data are processed in a random order $\rightarrow$ exchangeability.

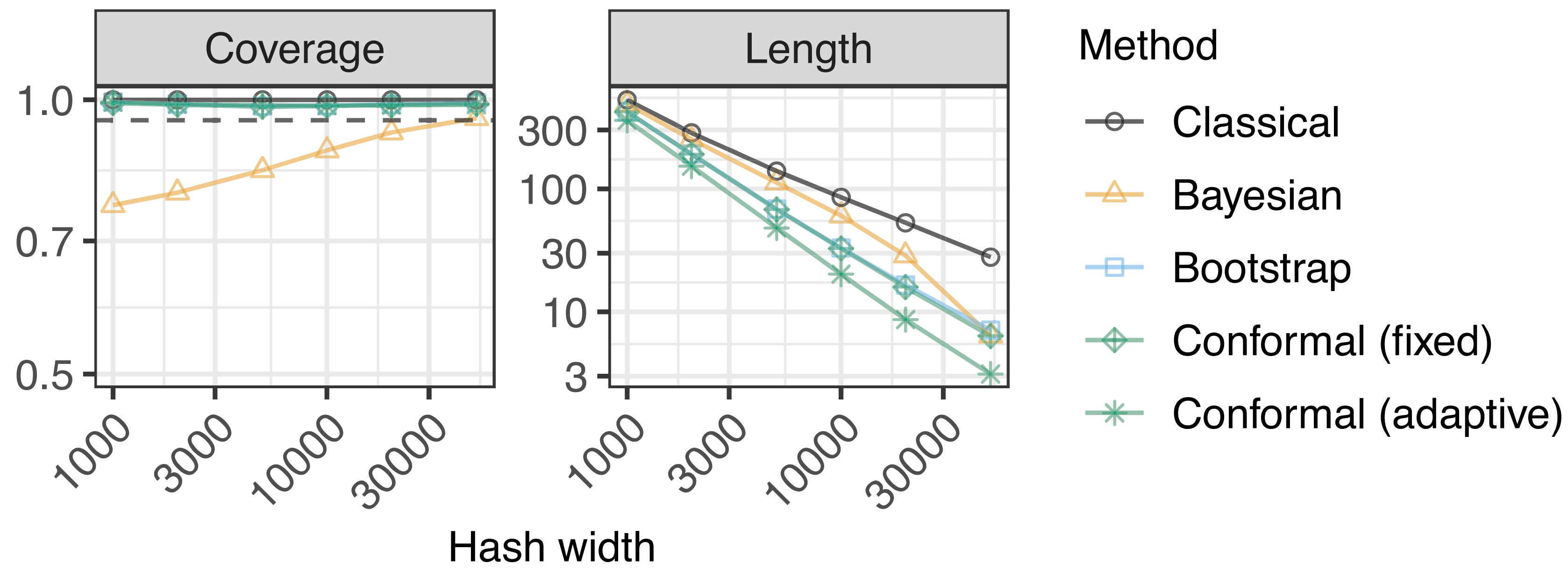
Analysis of 2-grams in English literature

Data: 18 open-domain pieces of classic English literature downloaded from the Gutenberg Corpus [Project Gutenberg, sent].

The goal is to count the frequencies of all *2-grams*—consecutive pairs of English words.

After some pre-proccessing, the number of 2-grams is $\approx$ 1,700,000.
The total number of all *possible* 2-grams is $\approx$ 650,000,000.

Sketch 1,000,000 2-grams, query 10,000 2-grams.
The data are processed in a random order $\rightarrow$ exchangeability.



Outline

1. Review of conformal prediction
2. Method: conformalized sketching
3. Numerical experiments
4. Discussion

Extensions (see papers)

Construction of conformal confidence intervals with:

- Frequency-range conditional coverage.

$$\mathbb{P}\left[f_n(Z_{n+1}) \in \hat{C}_{n,\alpha}(Z_{n+1}) \mid f_n(Z_{n+1}) \in B\right] \geq 1-\alpha,\ \forall B \in \mathcal{B}.$$

Extensions (see papers)

Construction of conformal confidence intervals with:

- Frequency-range conditional coverage.

$$\mathbb{P}\left[f_n(Z_{n+1}) \in \hat{C}_{n,\alpha}(Z_{n+1}) \mid f_n(Z_{n+1}) \in B\right] \geq 1 - \alpha, \ \forall B \in \mathcal{B}.$$

- Coverage for unique queries.

$$\begin{aligned} Z_1, \ldots, Z_n, Z_{n+1}, \ldots, Z_{n+M} &\overset{\text{exch.}}{\sim} P_Z, \\ Z^* &\sim \text{Uniform}\left[\textsc{Unique}(Z_{n+1}, \ldots, Z_{n+M})\right]. \\ \mathbb{P}\left[f_n(Z^*) \in \hat{C}_{n,\alpha}(Z^*)\right] &\geq 1 - \alpha. \end{aligned}$$

Extensions (see papers)

Construction of conformal confidence intervals with:

- Frequency-range conditional coverage.

$$\mathbb{P}\left[f_n(Z_{n+1}) \in \hat{C}_{n,\alpha}(Z_{n+1}) \mid f_n(Z_{n+1}) \in B\right] \geq 1 - \alpha, \ \forall B \in \mathcal{B}.$$

- Coverage for unique queries.

$$\begin{aligned} Z_1, \ldots, Z_n, Z_{n+1}, \ldots, Z_{n+M} &\overset{\text{exch.}}{\sim} P_Z, \\ Z^* \quad &\sim \text{Uniform}\left[\textsc{Unique}(Z_{n+1}, \ldots, Z_{n+M})\right]. \\ \mathbb{P}\left[f_n(Z^*) \in \hat{C}_{n,\alpha}(Z^*)\right] &\geq 1 - \alpha. \end{aligned}$$

- Robustness to distribution shift among the queries.

$$\begin{aligned} Z_1, \ldots, Z_n &\overset{\text{exch.}}{\sim} P_Z, \\ Z_{n+1} &\sim P'_Z. \end{aligned}$$

Conclusion

Conformalized sketching provides distribution-free inferences

- for any sketching algorithm
- for any (exchangeable) data set
- with valid marginal coverage (possibly also stronger coverage)

The key idea of data splitting is quite general and powerful: apply the statistical analysis (e.g., sketching) to some of the data, and use the rest of the data to track the performance.