Time-uniform, Computationally Efficient Post-selection Inference

Arun Kumar Kuchibhotla

17 Dec, 2024

Carnegie Mellon University

Joint work

This is a joint work with Siddhaarth Sarkar, CMU.

Table of contents

Introduction to Post-selection Inference

Inference: confidence intervals

- ⋆ Statistical inference is the cornerstone of statistics and is a necessary ingredient in any rigorous scientific study.
- ⋆ Traditional statistical inference deals with the inference for a functional $\theta(P), P \in \mathcal{P}$, when the functional is decided independently of the data.
- ⋆ For example, $\theta(P)$ could be the mean of P or a slope in a linear regression model.
- ⋆ In such a setting, assuming the existence of an estimator $\widehat{\theta}_n$ based on n observations such that

$$r_n(\widehat{\theta}_n - \theta(P)) \xrightarrow{d} L,$$

a confidence interval can be constructed as

$$\widehat{\mathrm{CI}}_{n,\alpha} := \left[\widehat{\theta}_n - \frac{\widehat{q}_{1-\alpha/2}}{r_n}, \widehat{\theta}_n + \frac{\widehat{q}_{\alpha/2}}{r_n}\right],$$

where $\widehat{q}_\gamma$ represents an estimate of the γ-th quantile of the random variable L.

Post-selection Inference

- ⋆ Unlike the setting of statistical theory, data analysts or stakeholders often take the functional to be data dependent.
- ⋆ This, most often, arises from a preliminary exploratory data analysis and then the analyst forms a "suitable" hypothesis to test.
- ⋆ Hence, we need inference for a data dependent functional $\widehat{\theta}(P)$.
- ⋆ For a concrete example, consider the data setting with one response Y and p covariates $X_1, \ldots, X_p$. The functionals one could consider are marginal correlations, i.e.,

  $$\theta_j(P) := \text{Corr}(Y, X_j) = \mathbb{E}[YX_j]. \quad \text{(assuming zero mean, unit var.)}$$

  These are data *independent* functionals. The analyst after performing univariate analysis might be interested in testing the hypothesis $H_0 : \theta_{\widehat{j}}(P) = 0$ where $\widehat{j}$ is the index of covariate that maximizes the correlation with Y in the data.

Post-selection Inference

- ⋆ The fundamental hurdle in post-selection inference is that $n^{1/2}(\widehat{\theta}_{\widehat{j}} - \theta_{\widehat{j}}(P))$ does not have a normal distribution, even asymptotically. Selection skews the estimator.
- ⋆ There is a rich literature on post-selection inference, and one of the proposed methods is simultaneous inference.
- ⋆ Simultaneous inference works by performing inference for all functionals that the analyst *could have* chosen.
- ⋆ In our example, if we know that the analyst will choose one of $\theta_j(P), 1 \leq j \leq p$ at the end of his/her exploration, we can report confidence intervals such that

$$\mathbb{P}\left(\bigcap_{j=1}^{p}\{\theta_j(P) \in \widehat{\mathrm{CI}}_{n,\alpha}^{(j)}\}\right) \geq 1-\alpha \quad \mathbb{R}\textit{ightarrow} \quad \mathbb{P}\left(\theta_{\widehat{j}}(P) \in \widehat{\mathrm{CI}}_{n,\alpha}^{(\widehat{j})}\right) \geq 1-\alpha.$$

No matter how $\widehat{j}$ is chosen.

Disadvantages of Simultaneous Inference

- ⋆ Although simultaneous inference gives a lot of flexibility in the analyst's selection method, it comes with certain disadvantages.
- ⋆ One has to specify the "universe" of selection.
- ⋆ Computation of simultaneous confidence intervals is "NP-hard" because one has to compute all the estimators in the universe for the construction of the confidence interval.
- ⋆ The validity of simultaneous confidence intervals is also restricted by the universe. The larger the universe, the more restrictive the conditions should be for validity.
- ⋆ In addition, simultaneous inference also cannot account for selection arising through sample size randomness.
- ⋆ In what follows, we discuss a simple framework to avoid specification of the universe, NP-hard computation, and restrictive assumptions on the data-generating process.

The Proposal

The Idea

- ⋆ Suppose that we are to obtain data from a distribution $P \in \mathcal{P}$. P is assumed to be supported on a subset of $\mathbb{R}^d$.
- ⋆ Suppose we can construct a data-dependent set of distributions $\widehat{\mathcal{P}}_{n,\alpha} \subseteq \mathcal{P}$ such that

$$\inf_{P \in \mathcal{P}} \mathbb{P}_P(P \in \widehat{\mathcal{P}}_{n,\alpha}) \geq 1 - \alpha, \tag{1}$$

then, for any functional $\theta : \mathcal{P} \to \mathbb{R}$, defining the set

$$\widehat{\mathrm{CI}}_{n,\alpha}(\theta) \ := \ \{\theta(P) : \ P \in \widehat{\mathcal{P}}_{n,\alpha}\},$$

we get

$$\inf_{P \in \mathcal{P}} \mathbb{P}_P \left(\theta(P) \in \widehat{\mathrm{CI}}_{n,\alpha}(\theta) \text{ for all functionals } \theta\right) \geq 1 - \alpha.$$

- ⋆ In particular, for *any* data-dependent functional $\widehat{\theta} : \mathcal{P} \to \mathbb{R}$, we get

$$\inf_{P \in \mathcal{P}} \mathbb{P}_P \left(\widehat{\theta}(P) \in \widehat{\mathrm{CI}}_{n,\alpha}(\widehat{\theta})\right) \geq 1 - \alpha.$$

- ⋆ Note that computation only involves the chosen functional $\widehat{\theta}$ and not the universe. Validity depends only on (2).

How is this any easier?

- ⋆ Construction of a data-dependent set of distributions $\widehat{\mathcal{P}}_{n,\alpha} \subseteq \mathcal{P}$ satisfying

$$\inf_{P \in \mathcal{P}} \mathbb{P}_P(P \in \widehat{\mathcal{P}}_{n,\alpha}) \geq 1 - \alpha, \tag{2}$$

  might look daunting.

- ⋆ To show that it is not, let us consider the one-dimensional case. 1-d distributions are characterized by the CDFs.

- ⋆ The classical DKW inequality implies

$$\mathbb{P}\left(\sup_x |\widehat{F}_n(x) - F_P(x)| \leq \sqrt{\frac{\log(2/\alpha)}{2n}}\right) \geq 1 - \alpha.$$

  Here $\widehat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}$ and $F_P(x) = \mathbb{P}_P(X \leq x)$.

- ⋆ Hence, an example of $\widehat{\mathcal{P}}_{n,\alpha}$ is the collection of distributions with CDFs lying between $\widehat{F}_n(x) - \sqrt{\log(2/\alpha)/2n}$ and $\widehat{F}_n(x) + \sqrt{\log(2/\alpha)/2n}$ for all x.

Impossibility Conflicts

- ⋆ While it is possible to construct confidence sets for distributions, it might not yield any useful confidence intervals for some functionals.
- ⋆ For example, given a DKW confidence set for CDF, we can construct an (almost) optimal confidence interval for the population median.
- ⋆ On the other hand, if we are interested in the mean, then the DKW confidence set yields the trivial confidence set of $\mathbb{R}$ for the mean.
- ⋆ This happens because no confidence set for the distribution can provide non-trivial information about the tails.
- ⋆ This can be escaped by restricting the collection of distributions $\mathcal{P}$.
- ⋆ For the mean example, Anderson (1969) considered the restriction of boundedness on $\mathcal{P}$. We consider more general conditions such as moment boundedness. (More about this in the following.)

What about the multivariate case?

- ⋆ In the one-dimensional case, the CDF is informative and sufficient enough for most functionals. In the multivariate case, the CDF is not enough.
- ⋆ As a generalization, for example, consider concentration inequalities for

$$\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in A\} - \mathbb{P}_P(X \in A) \right|, \tag{3}$$

  for a class of sets $\mathcal{A}$.
- ⋆ Moreover, in the 1-d case, we have distribution-free confidence sets. For example, (assuming continuity of $F_P(\cdot)$)

$$\sup_x |\widehat{F}_n(x) - F_P(x)| \stackrel{d}{=} \sup_{u \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq u\} - u \right|,$$

  where $U_1, \ldots, U_n$ are IID standard uniform random variables. This implies that one can construct (almost) exact confidence sets for P.
- ⋆ This distribution-free character is lost in the multivariate case, for computing bounds on (3).

Wald and Tukey Solution

⋆ Tukey, generalizing an idea of Wald, created statistically equivalent blocks $B_1, \ldots, B_{n+1}$ from multivariate data $X_1, \ldots, X_n$ such that

$$(\mu_P(B_1), \ldots, \mu_P(B_{n+1})) \stackrel{d}{=} (S_1, S_2, \ldots, S_n, S_{n+1}),$$

where $\mu_P(B) = \mathbb{P}_P(X \in B)$ and $S_j = U_{(j)} - U_{(j-1)}$ represent the spacings of standard uniform random variables.

⋆ Hence, we can construct a distribution-free confidence set for P by considering the known distribution of the spacings of uniform random variables.

⋆ An example of this construction is to cut the space $\mathbb{R}^d$ recursively based on different univariate projections of the data: Order data with respect to the first coordinate, split $\mathbb{R}^d$ into two parts based on the largest value of the first coordinate. Remove the observation with the largest first coordinate, repeat this with the second coordinate, and so on.

Statistically Equivalent Blocks

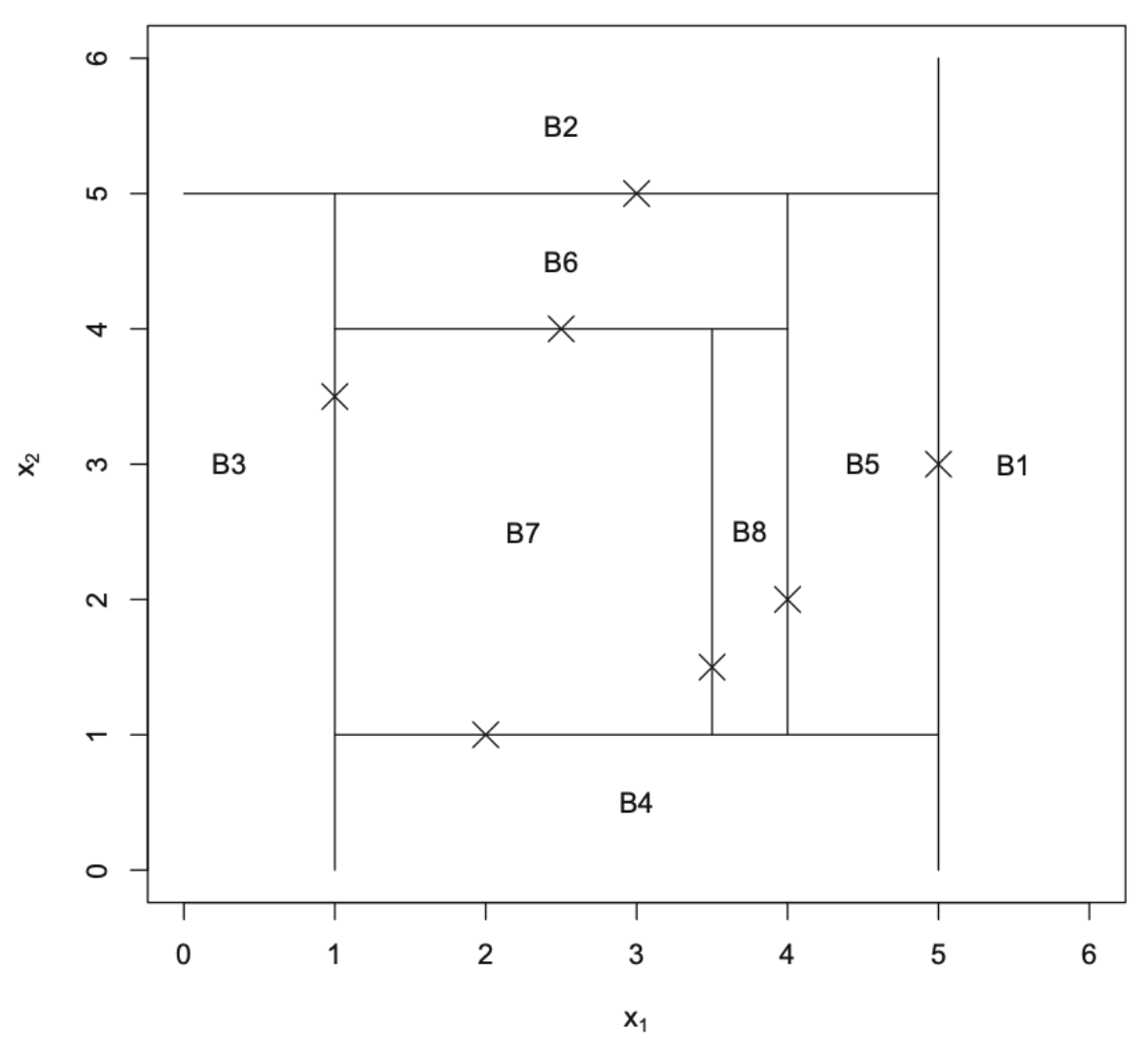


Figure 1: Statistically Equivalent Blocks: Illustration (Credit: Liu et al. (2022, Stat. in Med.))

Application to Population Mean

Confidence Intervals for Mean

- ⋆ Consider the special problem of constructing confidence intervals for the mean of a univariate distribution.
- ⋆ Although simple, it has far reaching applications and implications.
- ⋆ Note that

$$\mathbb{E}_P[X] = \int_0^1 F_P^{-1}(\delta)d\delta = \int_0^\infty (1 - F_P(x))dx + \int_{-\infty}^0 F_P(x)dx.$$

- ⋆ If we know $F_P(x) \in [\ell_\alpha(x), u_\alpha(x)]$ for all x with a probability of at least $1 - \alpha$, then computing bounds from above would yield $\mathbb{R}$.
- ⋆ Anderson (1969) considered random variables with support $[0, 1]$ to get a non-trivial confidence intervals.

Inference for Mean

⋆ We consider the general condition

$$\mathbb{E}_P[H(|X|)] \leq K, \tag{4}$$

for a non-negative, non-decreasing, even function $H : [0, \infty) \to [0, \infty)$.

⋆ We also assume $\lim_{|x|\to\infty} H(|x|)/|x| > 0$ so that (4) implies the existence of the mean.

Assumption	$H(x)$
Bounded r.v.	$\|x\|^\infty \mathbf{1}\{\|x\| > M\}$
Light tails	$\exp(x^2/t^2)$ for some $t \in \mathbb{R}$
Heavy tails	$\|x\|^k$ for any $k > 1$
Heavier tails	$\|x\| \log(\|x\|)$

⋆ Now, the confidence interval is

$$\widehat{\mathrm{CI}}_{n,\alpha} := \left[\inf_{F(x)\in[\ell_\alpha(x),u_\alpha(x)]\forall x,(4)} \int x dF(x), \sup_{F(x)\in[\ell_\alpha(x),u_\alpha(x)]\forall x,(4)} \int x dF(x) \right].$$

Width of the Confidence Interval

- ⋆ The width of the resulting confidence interval is heavily influenced by the choice of the confidence band and the constraint.
- ⋆ For example, with the DKW bound and constraint $\mathbb{E}[H(X/K)] \leq 1$, we get

$$\text{Width} \leq 4K\sqrt{\frac{\log(2/\alpha)}{2n}}H^{-1}\left(\sqrt{\frac{n}{4\log(2/\alpha)}}\right).$$

- ⋆ Hence, the width is $O(n^{-1/2})$ if and only if X is a bounded random variable. Even with sub-Gaussian random variables, the width is of the order $\sqrt{\log(n)/n}$. With $H(x) = x^2$, the width is of the order $n^{-1/4}$.
- ⋆ On the other hand, with confidence band of the type $\text{KL}(\widehat{F}_n(x), F(x)) \leq \kappa_\alpha$ for all x, we get

$$\text{Width} \leq K\kappa_\alpha\sqrt{\frac{\log\log n}{n}}, \quad \text{if} \quad H(x) = x^2.$$

- ⋆ The $\log\log n$ factor can also be removed if we use debiased KL confidence bands.

Computation and Simulations

Computation of the confidence intervals

- ⋆ Computation of the confidence interval requires finding the supremum and infimum of the integrals over a set of distribution functions.
- ⋆ This is in fact a linear programming problem in the space of probability measures.
- ⋆ The primal problem is

$$\sup / \inf \int x dF \text{ such that}$$

$$\ell_\alpha(x) \leq G(x) \leq u_\alpha(x) \quad \forall x \in \{X_1, \ldots, X_n\}$$

$$\int H(x) dG \leq K$$

$$\int dG = 1$$

G is a non-negative measure

Dual formulation

⋆ The dual is a linear semi-infinite programming (LSIP) problem.

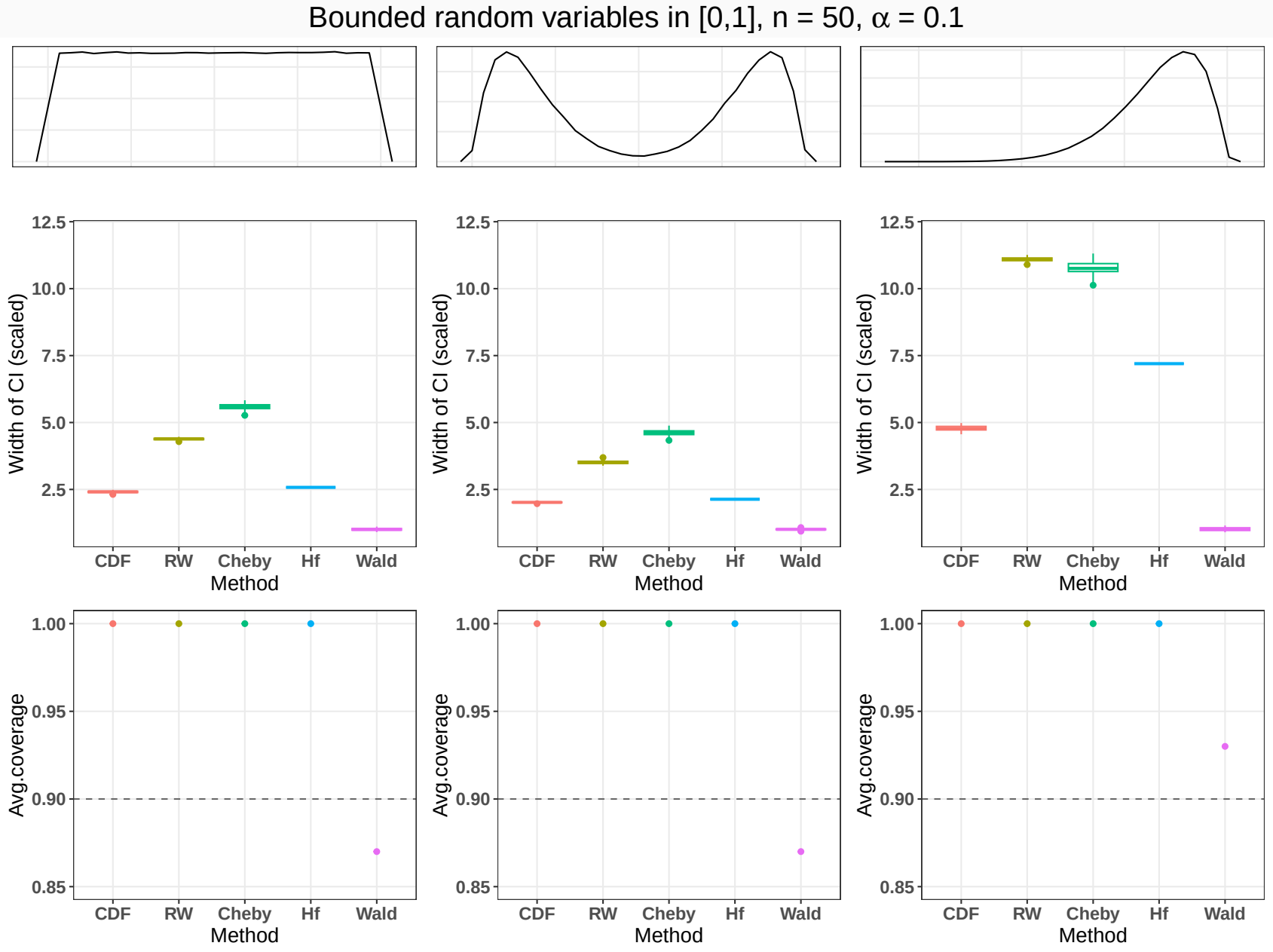
$$\sup / \inf \sum_{i=1}^{n} (\lambda_i^u u_\alpha(X_i) - \lambda_i^\ell \ell_\alpha(X_i)) + \lambda^H K + \lambda^p \text{ such that}$$

$$\sum_{i=1}^{n} (\lambda_i^u - \lambda_i^\ell) \mathbf{1}\{x \le X_i\} + \lambda^H H(x) + \lambda^p \ge x \quad \forall x \in \mathbb{R}$$

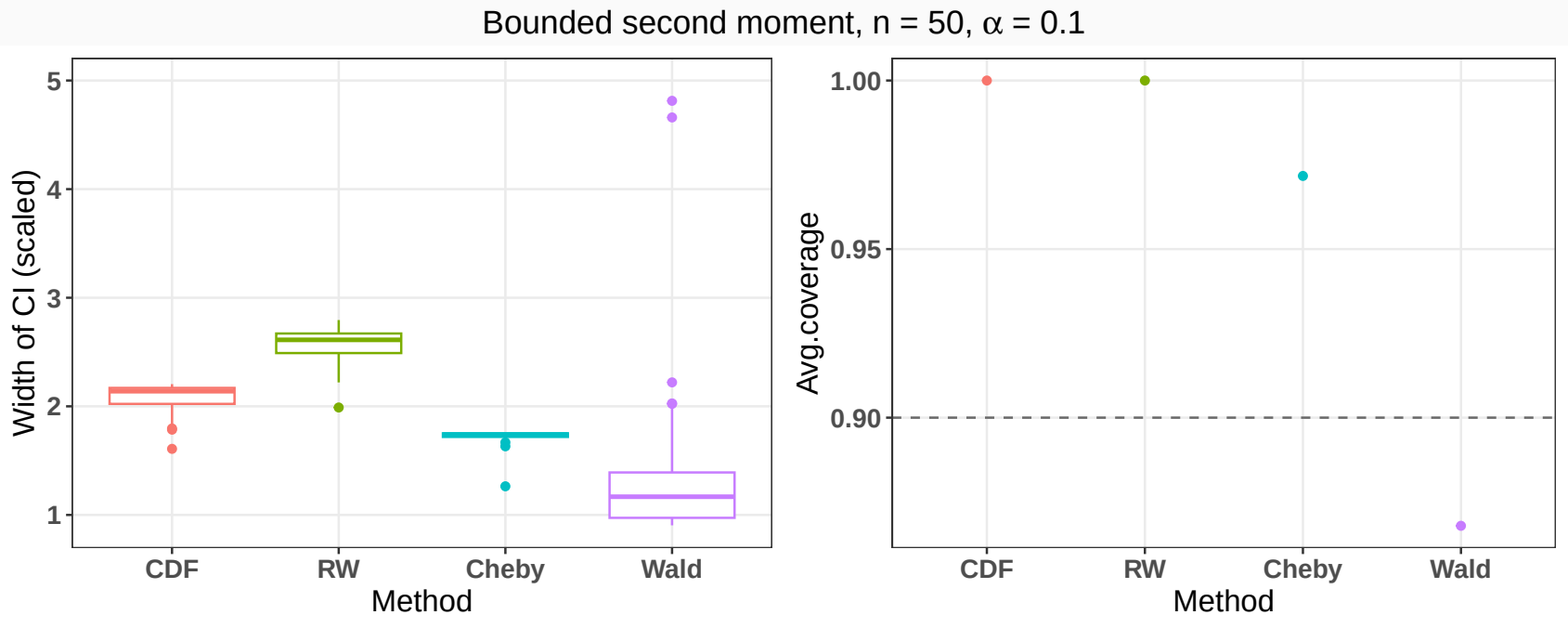
$$\lambda_I^u, \lambda_I^\ell, \lambda^H \ge 0, \quad \lambda^P \in \mathbb{R}$$

⋆ Solvable! (via discretization algorithm + proper initialization)

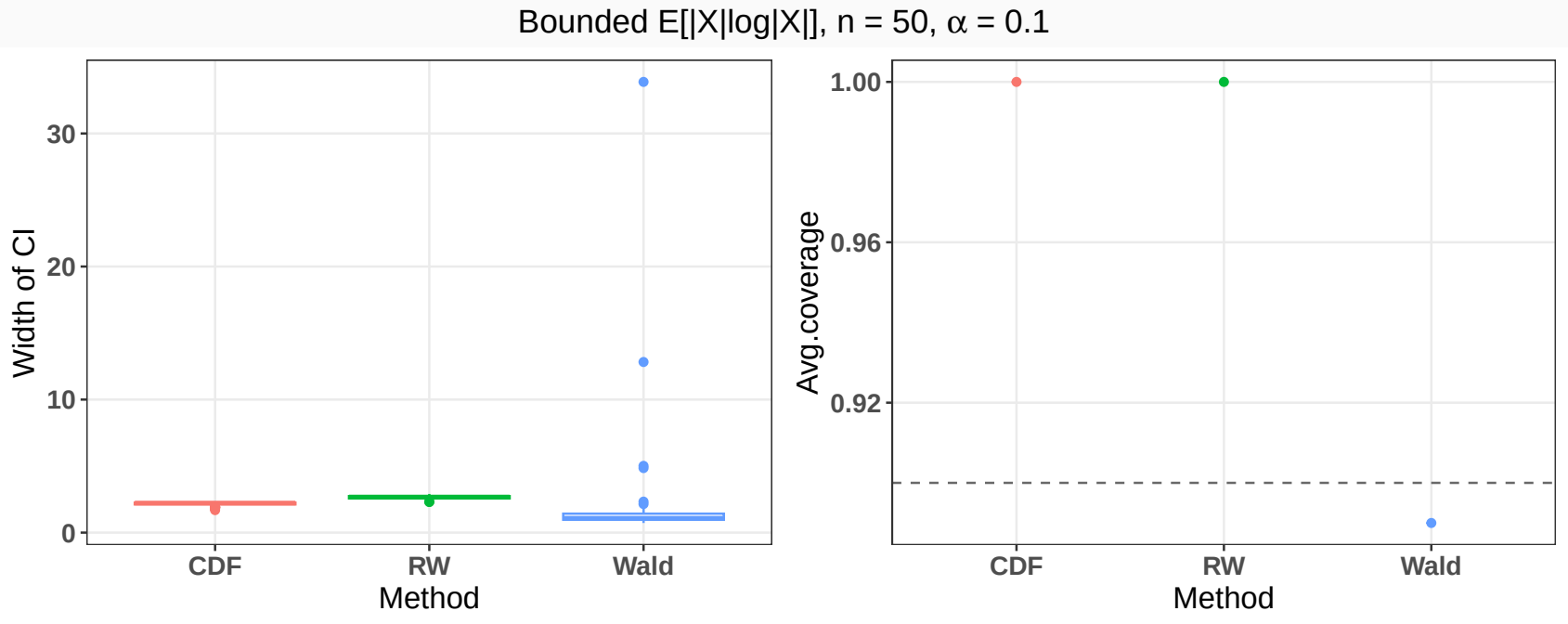
Simulations: Bounded r.v.



Simulations



Simulations

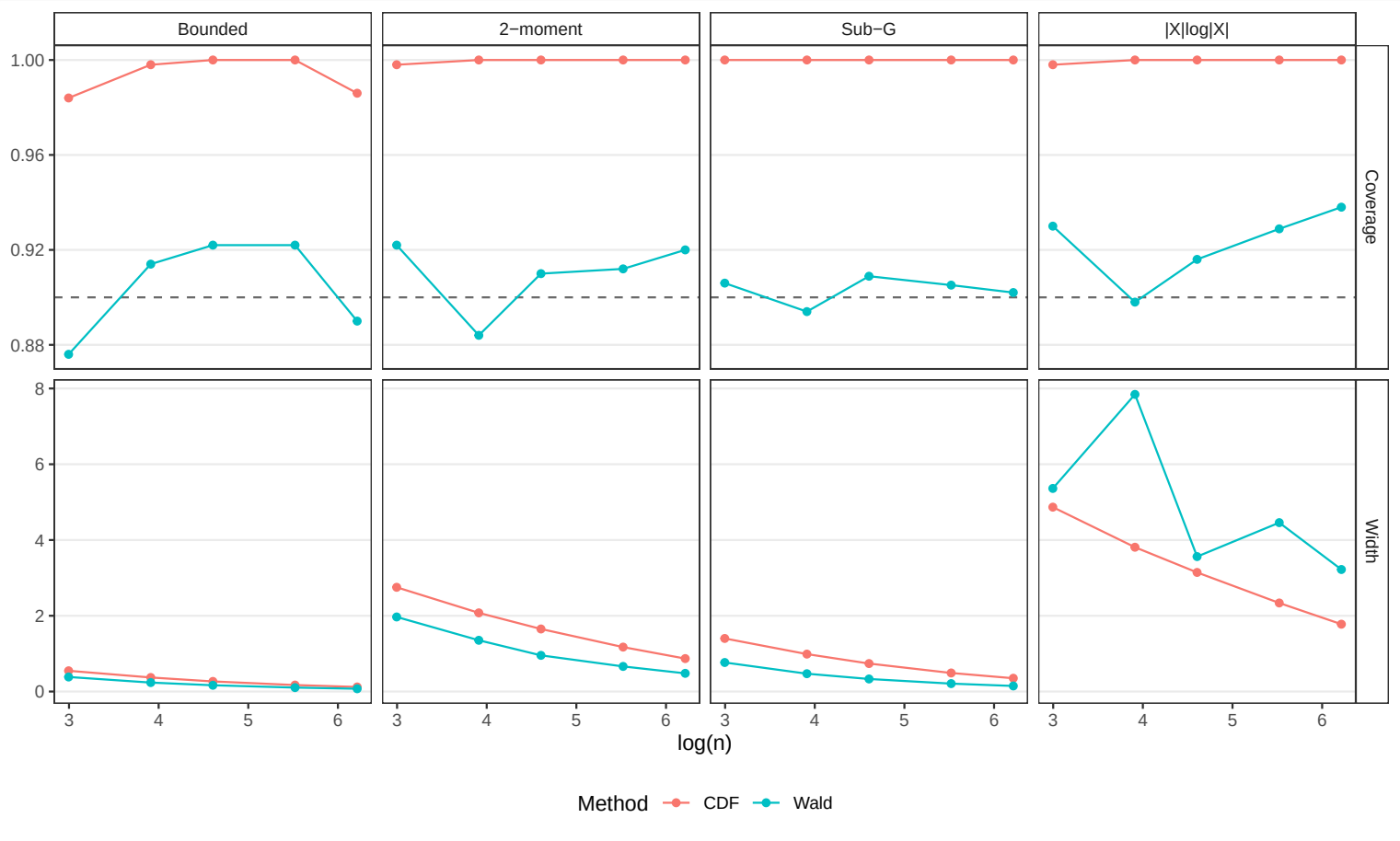


Simulations: growing sample size

Compare performance across different assumptions.

Data: $X \sim F_H$ such that

$$\mathbb{E}[H(X)] < \infty, \mathbb{E}\left[(H(X))^{1+\delta}\right] = \infty,$$



Conclusions

Conclusions

- ⋆ We have proposed a computationally efficient, assumption-lean post-selection valid confidence interval.
- ⋆ Time uniformity follows if we construct data-dependent classes of distributions such that

$$\inf_{P \in \mathbb{P}} \mathbb{P}_P \left(\bigcap_{n=1}^{\infty} \left\{ P \in \widehat{\mathcal{P}}_{n,\alpha} \right\} \right) \geq 1 - \alpha.$$

  This follows from Law of Iterated Logarithm (LIL) results for CDFs.
- ⋆ We have some preliminary results on the width of the confidence interval to show that they are a constant inflation of Wald intervals, when random variables have finite variance.
- ⋆ Much more to explore!!

Conclusions

- ⋆ We have proposed a computationally efficient, assumption-lean post-selection valid confidence interval.
- ⋆ Time uniformity follows if we construct data-dependent classes of distributions such that

$$\inf_{P \in \mathbb{P}} \mathbb{P}_P \left(\bigcap_{n=1}^{\infty} \left\{ P \in \widehat{\mathcal{P}}_{n,\alpha} \right\} \right) \geq 1 - \alpha.$$

  This follows from Law of Iterated Logarithm (LIL) results for CDFs.
- ⋆ We have some preliminary results on the width of the confidence interval to show that they are a constant inflation of Wald intervals, when random variables have finite variance.
- ⋆ Much more to explore!!

Thank You!