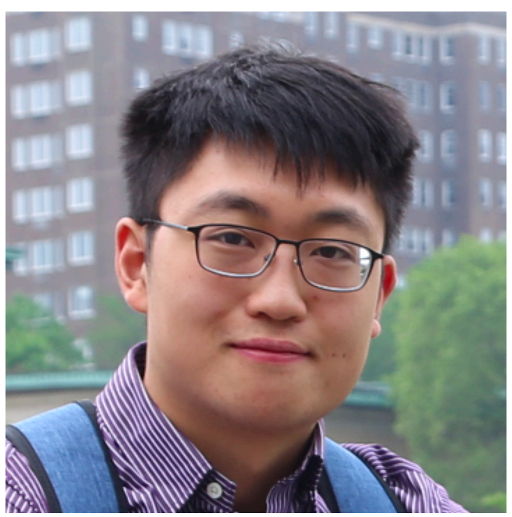
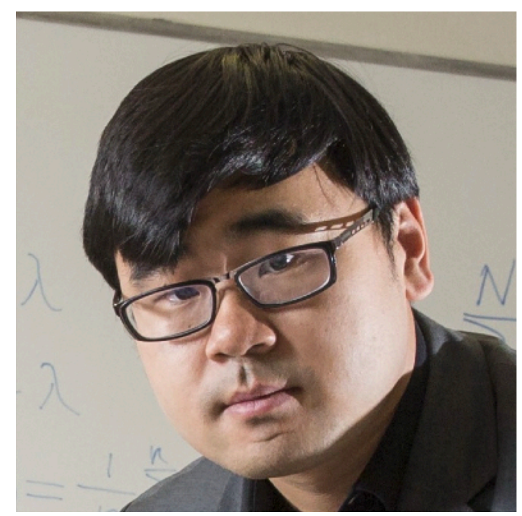
Post-selection inference with e-value based confidence intervals

International Seminar on Selective Inference
March 24, 2022

Joint work with:



Ziyu Xu (Neil)
(CMU)



Ruodu Wang
(Waterloo)

Aaditya Ramdas
(CMU)

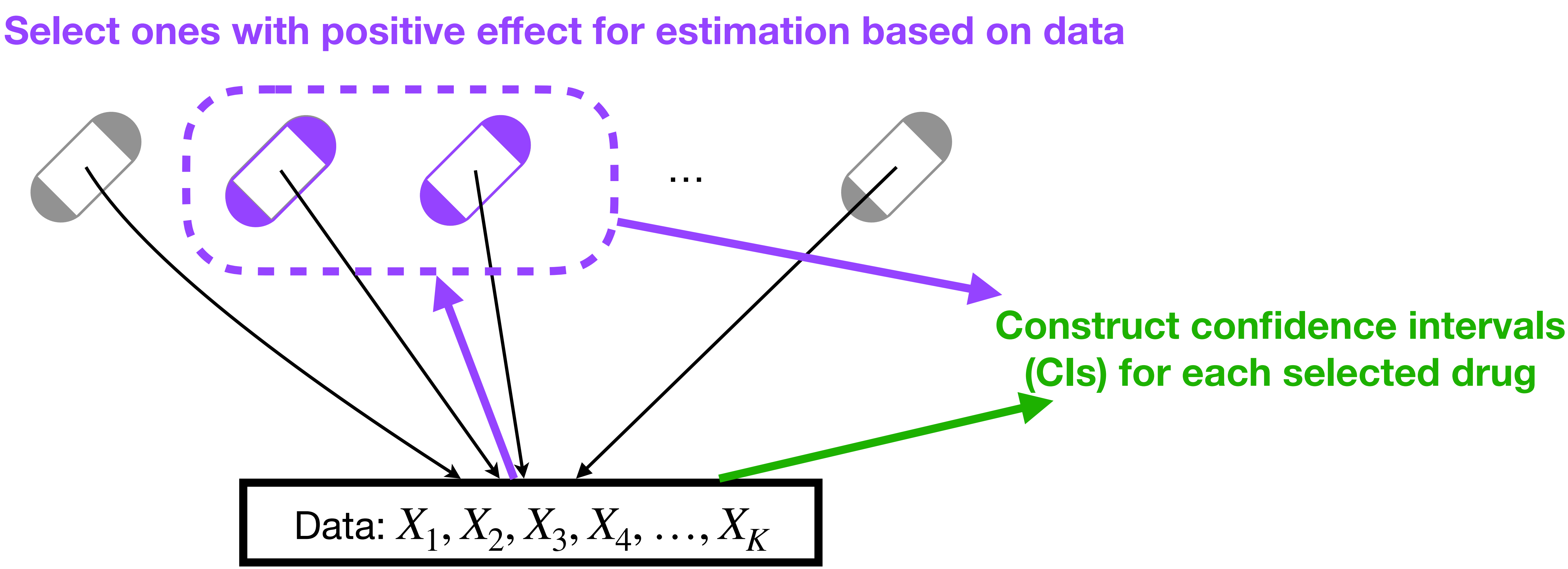
UNIVERSITY OF WATERLOO

arXiv: 2203.12572

Carnegie Mellon University

Motivating example: selecting most effective drug candidates

Initially, there are K drug candidates we wish to estimate the efficacy of.



Possible statistical guarantees under selection bias

$\theta_1, \ldots, \theta_K$ are the parameters we are initially interested in estimating.

We have: $C_i(\alpha)$ is the $(1-\alpha)$-CI we can construct for the ith parameter.

Marginal CI guarantee: $\mathbb{P}(\theta_i \in C_i(\alpha)) \geq 1-\alpha$

Selecting a subset $\mathcal{S}$ of parameters to estimate induces a selection bias.

We want: corrected levels $\alpha_1, \ldots, \alpha_K$ such that we can maintain some form of statistical validity.

False coverage rate (FCR): aggregate statistical validity

Define an empirical quantity: false coverage proportion (FCP).

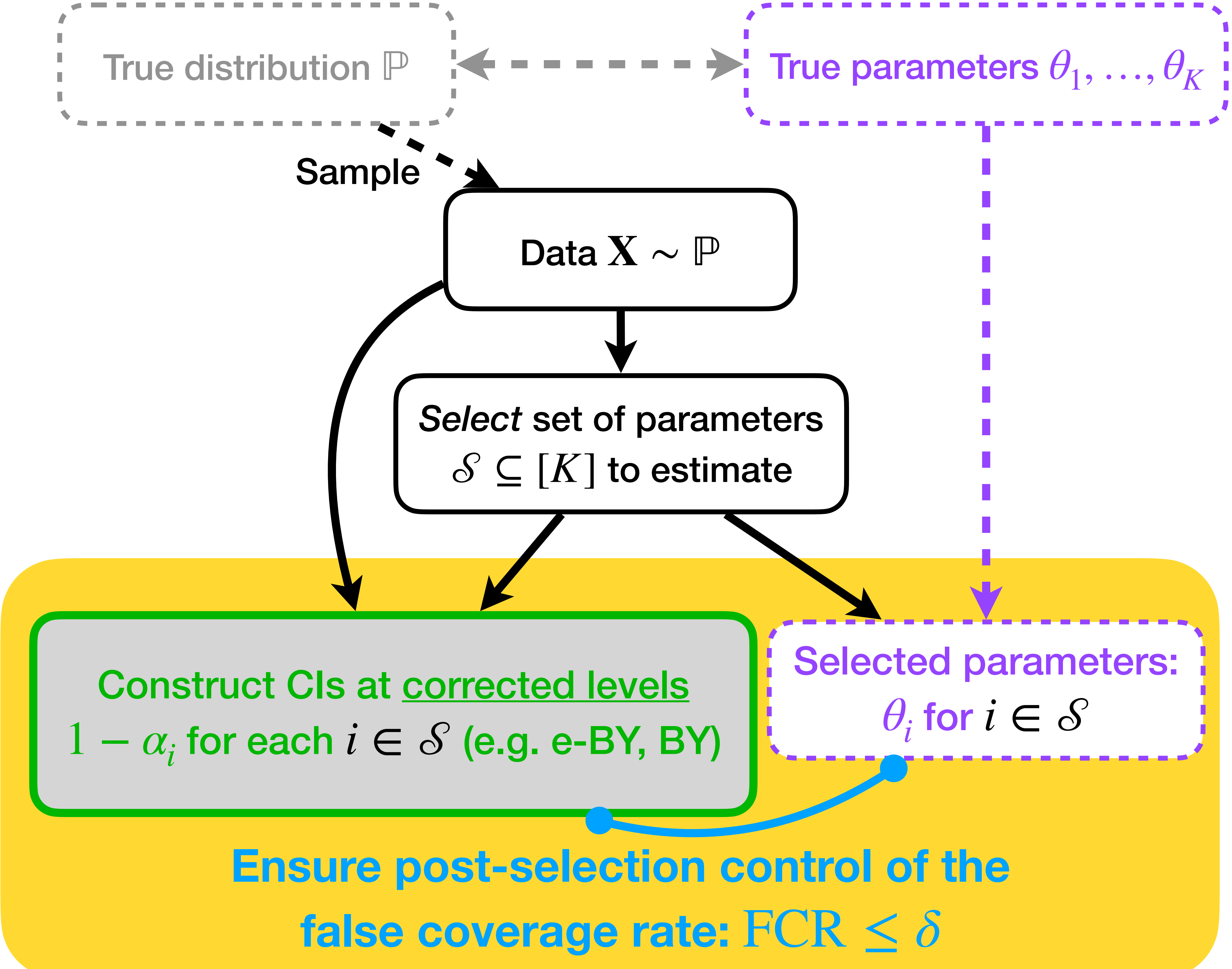
$$\mathrm{FCP} := \frac{\sum_{i \in \mathcal{S}} \mathbf{1}\{\theta_i \notin C_i(\alpha_i)\}}{|\mathcal{S}| \vee 1} \qquad \mathrm{FCR} := \mathbb{E}[\mathrm{FCP}]$$

FCR is an aggregate measure of false coverage across the CIs of selected parameters.

Analog of false discovery rate (FDR) from multiple testing

Benjamini and Yekutieli (2005) show how to control FCR with corrected marginal CIs

Post-selection inference with FCR guarantees



Current state-of-the-art: the BY procedure

Goal: Ensures $\mathrm{FCR} \leq \delta$

We have access to marginal CIs for each $i \in \{1,\ldots,K\}$:
$C_i(\alpha)$ s.t. $\mathbb{P}(\theta_i \in C_i(\alpha)) \geq 1 - \alpha$ for any $\alpha \in (0,1)$

In the **independent (or PRDS)** case: output $C_i\left(\frac{\delta R_i^{\min}}{K}\right)$ for each $i \in \mathcal{S}$

$1 \leq R_i^{\min} \leq |\mathcal{S}|$ is a value that depends on the selection rule

In the **dependent** case: output $C_i\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right)$ for each $i \in \mathcal{S}$

$\ell_K \approx \log K$ is the Kth harmonic number

Benjamini and Yekutieli (2005)

Calculating $R_i^{\min}$ requires knowledge of the selection rule.

Recall that $\mathbf{X} = (X_1, \ldots, X_K)$ is our sampled data.

$R_i^{\min} := \min \{ |\mathcal{S}(X_1, \ldots, x_i, \ldots, X_K)| : x_i \in \mathcal{X}_i$ and $i \in \mathcal{S}(X_1, \ldots, x_i, \ldots, X_K)\}$

Consider all possible $\mathcal{S}$ that could arise when both of the following are true:

1. X_i can be changed to any other possible data value x_i, but all other data $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$ remain fixed.
2. The ith parameter, θ_i, remains in the resulting selection set with changed x_i.

Many *known* selection rules achieve the upper bound of $|\mathcal{S}|$ e.g. CI above threshold, Benjamini-Hochberg (BH) etc.

For *unknown or ad-hoc* selection rules, no guarantees can be made i.e. cannot do better than $R_i^{\min} = 1$ (Bonferroni) and output $C_i\left(\frac{\delta}{K}\right)$.

Can also choose to fall back to dependent case guarantee: $C_i\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right)$.

Our method: the e-BY procedure

Goal: Ensures $\mathrm{FCR} \leq \delta$

We have access to **an e-CI** for each $i \in \{1, \ldots, K\}$.

Output $C_i\left(\dfrac{\delta |\mathcal{S}|}{K}\right)$ for each $i \in \mathcal{S}$

1. There is no reliance on selection rule (through $R_i^{\min}$) or change based on dependence structure.

2. E-BY requires access to e-CIs, a special class of CIs.

Head-to-head comparison of e-BY vs. BY

	e-BY	BY
Knowledge of selection rule	None required	Needed for independent/PRDS case through $R_i^{\min}$
Effect of dependence structure	No effect - always output $C_i\left(\frac{\delta\lvert\mathcal{S}\rvert}{K}\right)$	Independence/PRDS: $C_i\left(\frac{\delta R_i^{\min}}{K}\right)$ Dependent: $C_i\left(\frac{\delta\lvert\mathcal{S}\rvert}{K\ell_K}\right)$
Type of CI	Only e-CIs	All CIs

The BY procedure is a special case of the e-BY procedure obtained by calibrating CIs to e-CIs

Eliciting e-CIs easily from evidence

- Definition of an e-CI (and e-values)
- E-CIs from universal inference and supermartingales
- Calibration CIs to e-CIs (BY is a special case of e-BY)
- Simulations in a nonparametric setting

E-value: e is for expectation (is bounded by 1)

E is an **e-value** w.r.t. to a parameter θ if and only if:

1. E is nonnegative under θ, and
2. $\mathbb{E}_\theta[E] \leq 1$ under θ

E-values are analogs of p-values that have been extensively studied in recent work in testing and estimation (Shafer, Vovk, Grünwald, and others)

Fact: Under θ, $\mathbb{P}\left(E \geq \frac{1}{\alpha}\right) \leq \alpha$ for any $\alpha \in (0,1)$.

True by Markov's inequality!

Consequently, $1/E$ is a p-value.

Inverting e-values produces an e-CI

Denote the universe of parameters as Θ.

C is an **e-CI** if and only if there exists a family of e-values $E(\theta)$ where:

$$C(\alpha) := \left\{ \theta \in \Theta : E(\theta) < \frac{1}{\alpha} \right\}$$

Fact: Every e-CI C is a valid confidence interval (CI).

Proof: $\mathbb{P}(\theta \notin C(\alpha)) = \mathbb{P}\left(E(\theta) \geq \frac{1}{\alpha} \right) \leq \alpha$

This is by Markov's inequality, again

Proof of FCR control of e-BY

Recall e-BY outputs $C_i\left(\frac{\delta|\mathcal{S}|}{K}\right)$ for each $i \in \mathcal{S}$ where C_i is an e-CI for θ_i.

Proof that $\mathrm{FCR} \leq \delta$ for e-BY:

$$\mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\left\{\theta_i \notin C_i\left(\frac{\delta|\mathcal{S}|}{K}\right)\right\}}{|\mathcal{S}| \vee 1}\right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\left\{E_i(\theta_i) \geq \frac{K}{\delta|\mathcal{S}|}\right\}}{|\mathcal{S}| \vee 1}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{1}\left\{E_i(\theta_i)\delta|\mathcal{S}|/K \geq 1\right\}}{|\mathcal{S}| \vee 1}\right] \leq \sum_{i=1}^{K} \mathbb{E}\left[\frac{E_i(\theta_i)\delta|\mathcal{S}|/K}{|\mathcal{S}| \vee 1}\right]$$

($\mathbf{1}\{x \geq 1\} \leq x$ and $\mathcal{S} \subseteq \{1, \ldots, K\}$)

$$\leq \frac{\delta}{K}\sum_{i=1}^{K} \mathbb{E}\left[E_i(\theta_i)\frac{|\mathcal{S}|}{|\mathcal{S}| \vee 1}\right] \leq \delta$$

(def. of e-value)

Universal inference e-value

Universal inference (Wasserman, Ramdas, Balakrishnan 2020) is a method for deriving e-values/e-CIs whenever the likelihood function is known.

Receive i.i.d. data $A_1, \ldots, A_n$ and split equally into two datasets D_0, D_1

For any $\theta \in \Theta$, perform a likelihood ratio test between:

$H_0 : \theta$ is the true parameter, $H_1 : \theta$ is not the true parameter

Estimate any likelihood $\widehat{p}_1$ using D_1 (alternative likelihood)

Universal inference e-value: $E^{\mathrm{UI}}(\theta) := \dfrac{\widehat{p}_1(D_0)}{\boxed{\max_{p \in P_\theta} p(D_0)}}$

p: maximum likelihood of D_0 under null (set of distributions with parameter θ)

Universal inference e-value

Let p^* be the likelihood of the true distribution, and θ be the true parameter.

Proof $E^{\mathrm{UI}}(\theta)$ is an e-value:

$$\mathbb{E}[E^{\mathrm{UI}}(\theta) \mid D_1] = \mathbb{E}\left[\frac{\widehat{p}_1(D_0)}{\max_{p \in P_\theta} p(D_0)} \mid D_1\right] \leq \mathbb{E}\left[\frac{\widehat{p}_1(D_0)}{p^*(D_0)} \mid D_1\right]$$

(p^* has parameter θ)

$$= \int \frac{\prod_{i=1}^{n/2} \widehat{p}_1(A_i)}{\prod_{i=1}^{n/2} p^*(A_i)} \cdot \prod_{i=1}^{n/2} p^*(A_i)\ dA_1, \ldots, dA_n$$

$$= \int \prod_{i=1}^{n/2} \widehat{p}_1(A_i)\ dA_1, \ldots, dA_n = 1$$

Thus, $\mathbb{E}[\mathbb{E}[E^{\mathrm{UI}}(\theta) \mid D_1]] \leq 1$

Universal inference e-value

Define the universal inference e-CI:

$$C^{\mathrm{UI}}(\alpha) := \left\{ \theta \in \Theta : E^{\mathrm{UI}}(\theta) < \frac{1}{\alpha} \right\} = \left\{ \theta \in \Theta : \alpha \widehat{p}_1(D_0) < \max_{p \in P_\theta} p(D_0) \right\}.$$

WRB20 prove that $\mathbb{P}(\theta \in C^{\mathrm{UI}}(\alpha)) \geq 1 - \alpha$.

We can use universal inference to:

- estimate number of components in GMM in high dimensions
- estimate sparsity of regression problem
- determine if a distribution satisfy certain shape-constraints
- estimate parameters whenever we have likelihoods

Confidence sequences and e-CIs

In the sequential regime, samples come one at a time in a stream $A_1, A_2, \ldots$

An $(1-\alpha)$-**confidence sequence** is a sequence of intervals $(C^t(\alpha))_t$ where $\mathbb{P}(\forall t \in \mathbb{N} : \theta \in C^t(\alpha)) \geq 1-\alpha$

Example (Howard et al. 2021): If A_i are 1-sub-Gaussian,

$$C^t(\alpha) := \frac{1}{t}\sum_{i=1}^{t} A_i \pm \sqrt{\frac{\log\log 2t + 0.72\log(10.4/\alpha)}{t}}$$

is a $(1-\alpha)$-confidence sequence for estimating $\theta = \mathbb{E}[A_i]$.

Confidence sequences and e-CIs

Define a filtration $(\mathscr{F}_t)$ where $\mathscr{F}_t = \sigma(A_1, \ldots, A_t)$

A **stopping time** τ w.r.t. a filtration $(\mathscr{F}_t)$ is a random time that can determine whether it stops a time t based on the info in $\mathscr{F}_t$

$C^\tau(\alpha)$ is an $(1-\alpha)$-e-CI for any stopping time τ
and confidence sequence $(C^t(\alpha)_t)$

"Proof":

Confidence sequences are constructed by inverting nonnegative supermartingales.

Nonnegative supermartingales (that have initial value less than 1) are e-values at stopping times.

Confidence sequences and e-CIs

We can build sequential e-CIs for any situation where we have Chernoff bounds (Howard et al. 2020, 2021)

We can also build batch e-CIs from batch Chernoff bounds (Hoeffding, Bernstein, etc.)

We can also extend universal inference to the sequential regime.

Sequential e-CIs based off of nonnegative martingales are admissible (Ramdas et al. 2020)

Calibration: going from CI to e-CI

We can always *calibrate* a CI into an e-CI.

Based line of work about calibrating p-values into e-values (Shafer, Vovk, Wang, etc.)

A **calibrator** is a upper semicontinuous, nonincreasing function $f : [0,1] \times [0,\infty]$ such that:

$$\int_0^1 f(x)\ dx = 1$$

Let C be a CI. Every CI has an associated implicit p-value:

$$P^{\mathrm{dual}}(\theta) := \inf\ \{\alpha \in [0,1] : \theta \notin C(\alpha)\}$$

Consequently, define an e-value $E^{\mathrm{cal}}(\theta) := f(P^{\mathrm{dual}}(\theta))$

Calibration: going from CI to e-CI

Calibrated e-CI:

$$C^{\text{cal}}(\alpha) := \left\{\theta \in \Theta : E^{\text{cal}}(\alpha) < \frac{1}{\alpha}\right\} = C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right)$$

where $f^{-1}(x) = \sup\ \{p : f(p) \geq x\}$

Examples of calibrators:

- All or nothing: $f(p) = \dfrac{1}{\beta}\mathbf{1}\{p \leq \beta\}$ for any $\beta \in (0,1)$
- Power: $f(p) = \kappa p^{\kappa-1}$ for any $\kappa \in (0,1)$

Calibration implies BY = e-BY under dependence

The BY(δ, K) calibrator:

$$f^{\mathrm{BY}(\delta,K)}(p) = \frac{K}{\delta} \cdot \frac{1}{\lceil K\ell_K p/\delta \rceil}$$

With CI C_i, recall that the BY procedure outputs for each $i \in \mathcal{S}$:

$$C_i\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right)$$

With the e-CI C_i^{cal} calibrated with $f^{\mathrm{BY}(\delta,K)}$ from C_i, e-BY outputs:

$$C_i^{\mathrm{cal}}\left(\frac{\delta|\mathcal{S}|}{K}\right) = C_i^{\mathrm{cal}}\left(f^{\mathrm{BY}(\delta,K)^{-1}}\left(\frac{K\ell_K}{\delta|\mathcal{S}|}\right)\right) = C_i\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right)$$

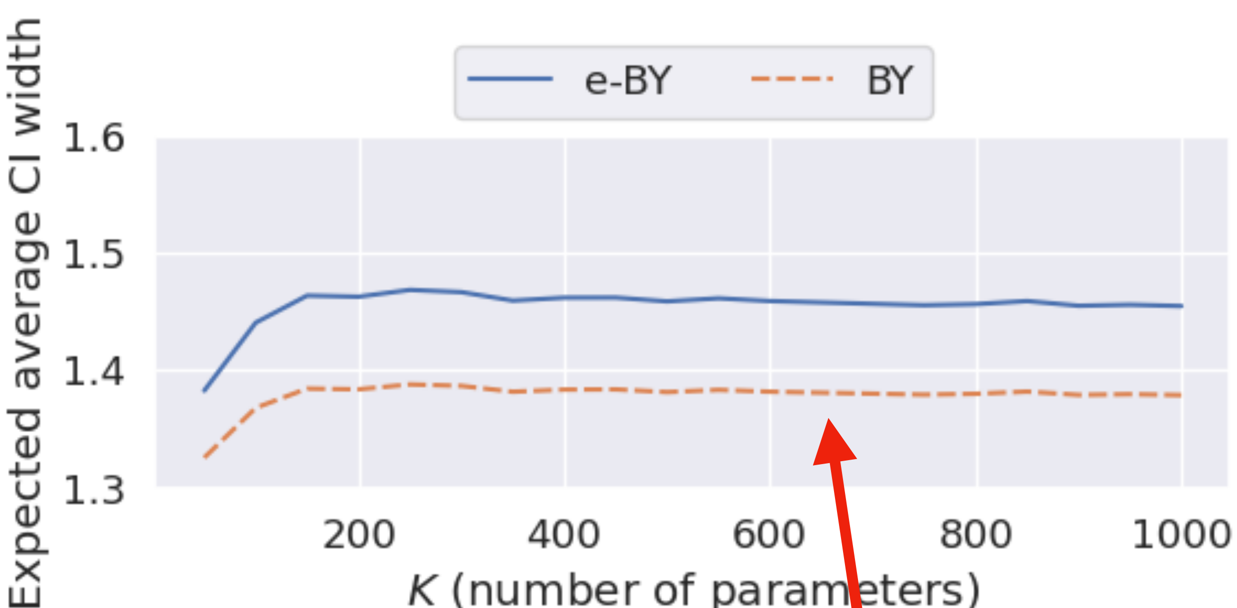
The BY procedure is a special case of e-BY.

Simulations for bounded random variables indicate e-BY is tighter under dependence

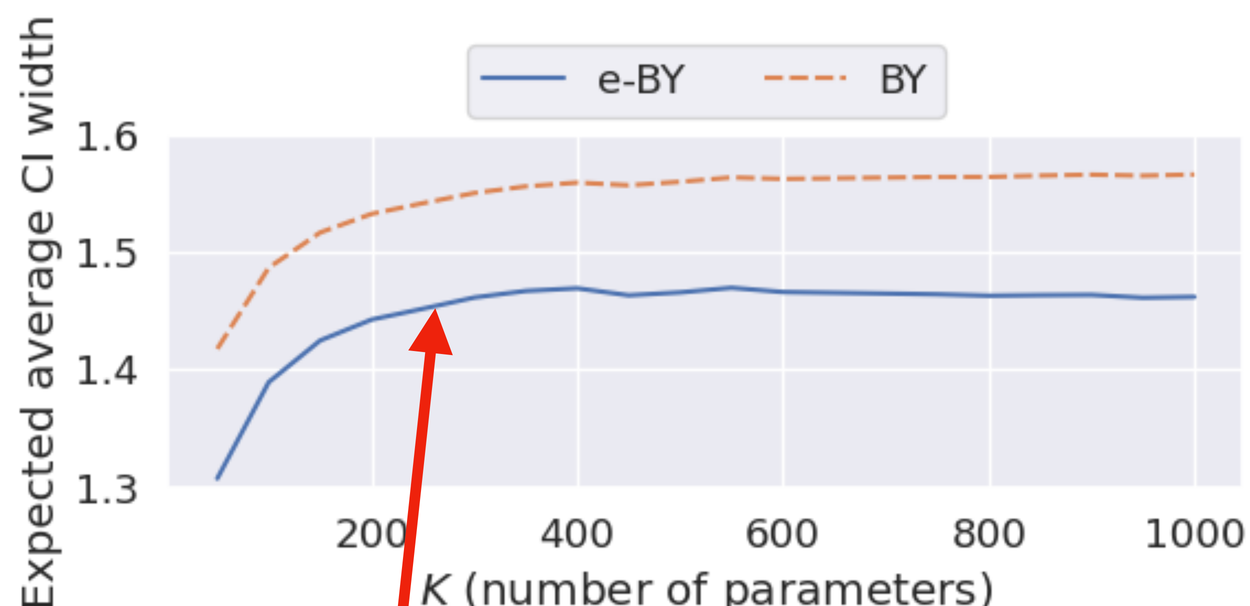
Nonparametric setting: estimate the mean of bounded random variables in $[-1,1]$

Hoeffding based CI for the BY procedure and e-CI for the e-BY procedure.

Select parameters that have solely positive $(1-\delta)$-CIs $\delta = 0.1$



Independent



Dependent

BY tighter than e-BY

e-BY tighter than BY

Takeaways

1. e-BY procedure provides FCR control with no assumptions about dependence and the selection rule, as opposed to the BY procedure.

2. e-BY operates only on a restricted class of CIs: e-CIs.

3. BY under dependence is a special case of e-BY.

4. e-CIs can are already used in many settings e.g. universal inference, sequential settings, Chernoff methods.

5. e-CIs are particularly tight in the sequential regime.

Thanks! **arXiv: 2203.12572**