Boosting e-BH via Conditional Calibration

Junu Lee

Joint work with Zhimei Ren

University of Pennsylvania, Department of Statistics and Data Science

International Seminar on Selective Inference, June 2024

Structure of the talk

- E-values: introduction, background, multiple testing
- e-BH-CC: Boosting e-BH via conditional calibration
- Three specific instantiations of e-BH-CC
  - implementation and simulation results

E-value: an alternative to the p-value

Testing the null hypothesis H_0:

- E-value e is the realization of an e-variable E:

$$E \geq 0,\ \mathbb{E}_{H_0}[E] \leq 1$$

- Reject H_0 when $e \geq 1/\alpha \Rightarrow$ level-α test

[Shafer '19; Grünwald et al. '24; Wang and Vovk '21 ...]

- P-value p is the realization of an p-variable P:

$$P \in [0,1],\ \mathbb{P}_{H_0}(P \leq t) \leq t, t \in (0,1)$$

- Reject H_0 when $p \leq \alpha \Rightarrow$ level-α test

Some nice properties of e-values

- $e_1,\ e_2$ are e-values $\Longrightarrow \dfrac{1}{2}(e_1 + e_2)$ is an e-value
- $e_1,\ e_2$ are e-values $\Longrightarrow e_1 e_2$ is an e-value if $\mathbb{E}_{H_0}[e_2 \mid e_1] \leq 1$

What are e-values?

- Likelihood ratio

  ex. $$\frac{d\mathcal{N}(\mu,1)}{d\mathcal{N}(0,1)}(z) = \exp(\mu z - \mu^2/2)$$

- Betting scores
- Bayes factors
- (Stopped) supermartingales
- ...

Connection between p-values and e-values

- If e is an e-value, $1/e$ is a p-value

  $$\mathbb{P}_{H_0}(1/e \leq t) = \mathbb{P}_{H_0}(e \geq 1/t) \leq t\mathbb{E}_{H_0}[e] \leq t,\ t \in (0,1)$$

- A p-value p can be transformed into an e-value through a calibrator f, defined as a non-increasing function satisfying

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

  e.g., $f(x) = \lambda x^{\lambda-1}, \lambda \in (0,1)$ [Wang and Vovk '21]

Testing multiple hypotheses

m null hypotheses: $H_1, H_2, \ldots, H_m$

H_j can be:

- Whether genetic variant *j* is associated with the phenotype of interest
- Whether gene *j* is differentially expressed in the treatment and control environment
- Whether bandit arm *j* has mean reward higher than some threshold r_0
- …

Goal: obtain a rejection set $R \subseteq \{1, \ldots, m\}$ while controlling the false discovery rate (FDR):

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{j\ \text{null}} \mathbf{1}\{j \in R\}}{\max(|R|, 1)}\right]$$

[Benjamini and Hochberg '95]

Multiple testing with FDR control

- Associate each H_j with a p-value p_j
- Obtain rejection set $R(p_1, \ldots, p_m)$
- The Benjanimi-Hochberg (BH) procedure
  - Provably controls the FDR if the p-values are independent or positively correlated
- Other variants w/ inflated or asymptotic control

[Benjamini and Yekutieli '01; Genovese and Wasserman '04; Storey et al. '04; Ferreira and Zwinderman '06; Farcomeni '07 ...]

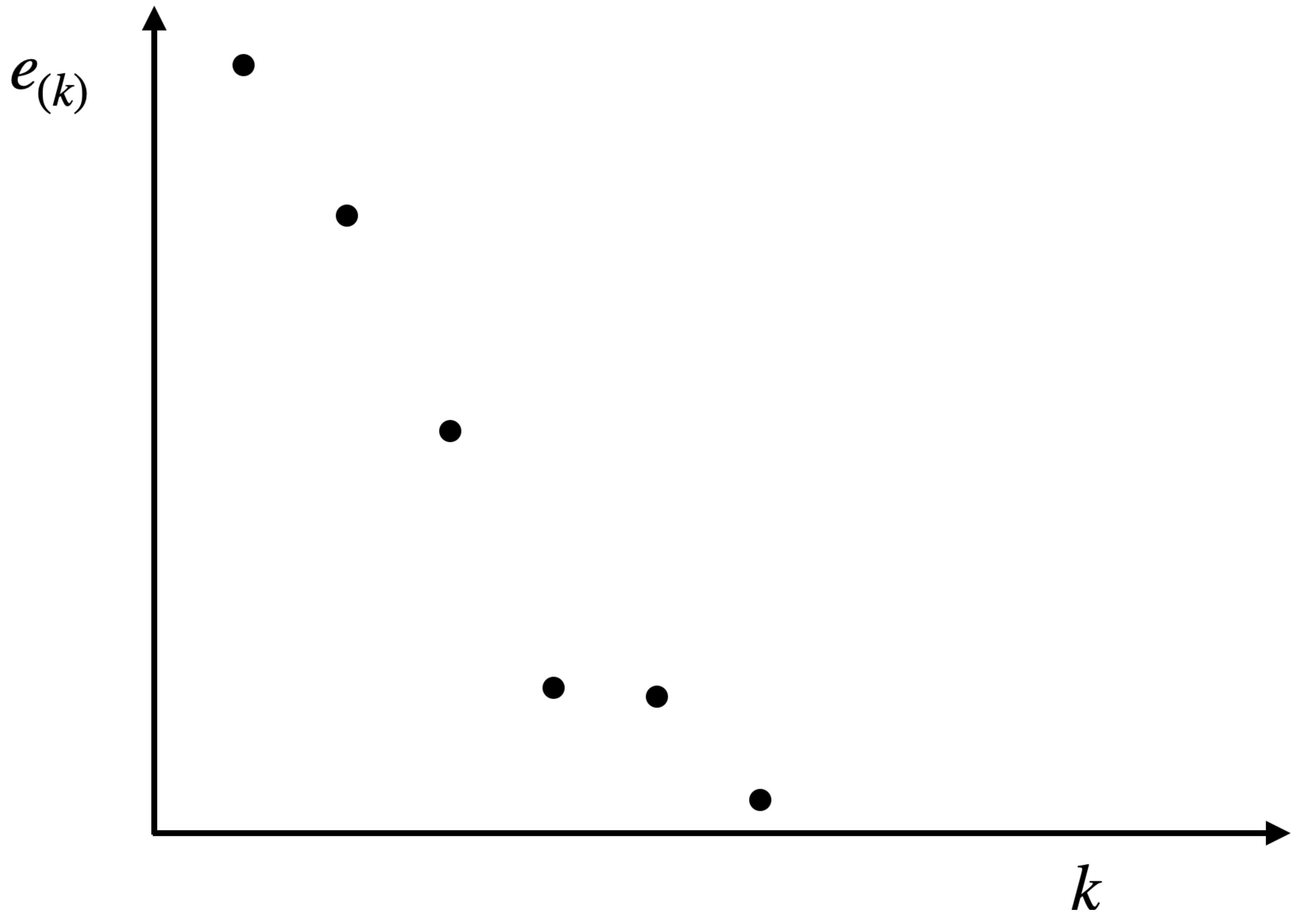
- Associate each H_j with an e-value e_j
- Obtain rejection set $R(e_1, \ldots, e_m)$
- The e-BH procedure
  - Provably controls the FDR under arbitrary dependence structure

[Wang and Ramdas '22]

The e-BH procedure

- A set of e-values $(e_1, \ldots, e_m)$
- Rank them in descending order:

$$e_{(1)} \geq \ldots \geq e_{(m)}$$



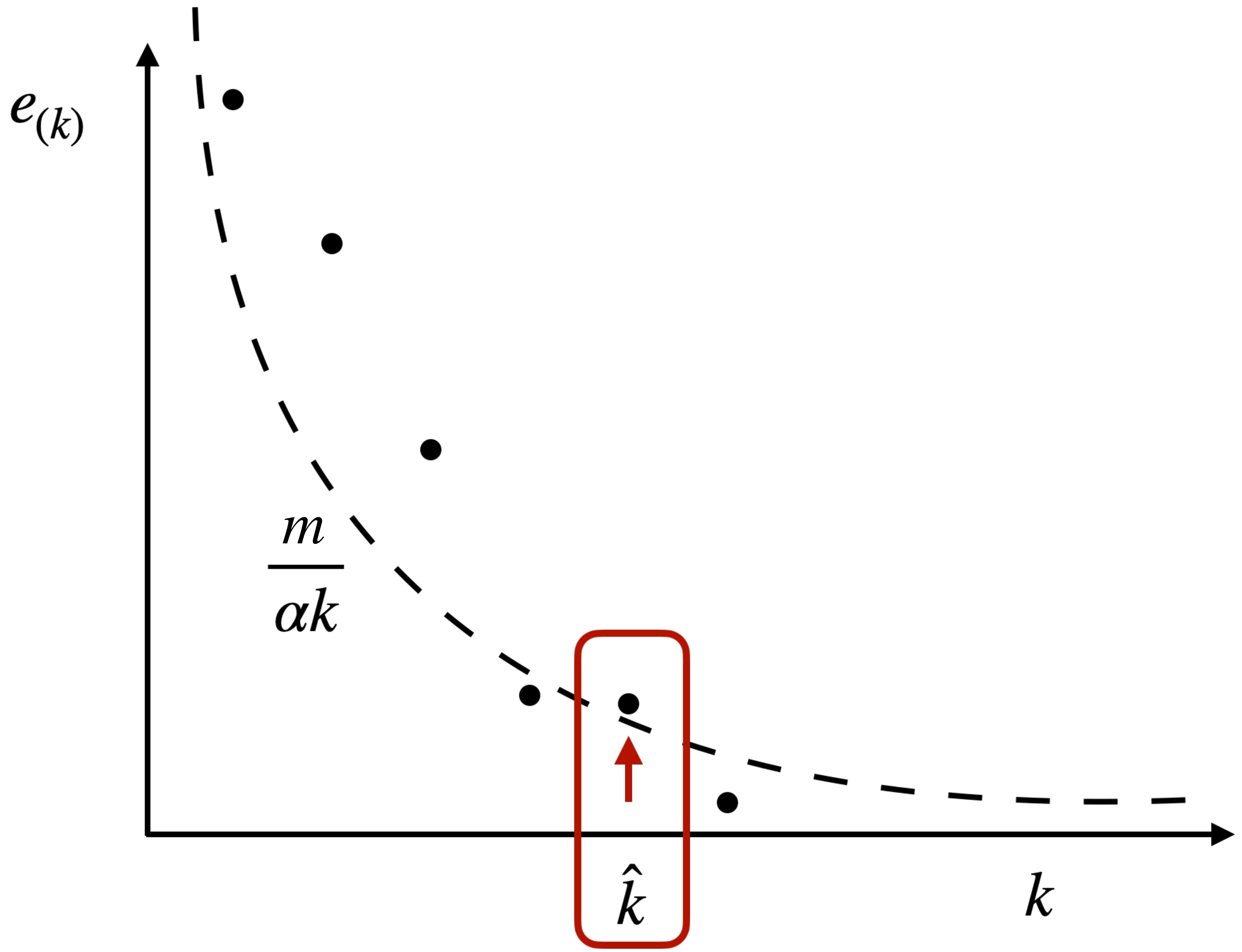
The e-BH procedure

- A set of e-values $(e_1, \ldots, e_m)$
- Rank them in descending order:

$$e_{(1)} \geq \ldots \geq e_{(m)}$$

- Reject the $\hat{k}$ largest e-values, where

$$\hat{k} = \max \left\{ k \in [m] : e_{(k)} \geq \frac{m}{\alpha k} \right\}$$



Why the FDR control?

$$j \in R \iff e_j \geq \frac{m}{\alpha |R|}$$ (self-consistency of e-BH)

$$\text{FDR} = \sum_{j \text{ null}} \mathbb{E}\left[\frac{\mathbf{1}\{j \in R\}}{\max(|R|,1)}\right] = \sum_{j \text{ null}} \mathbb{E}\left[\frac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)}\right]$$

for $t > 0$, $\mathbf{1}\{X \geq t\} \leq \frac{X}{t}$ $\longrightarrow$

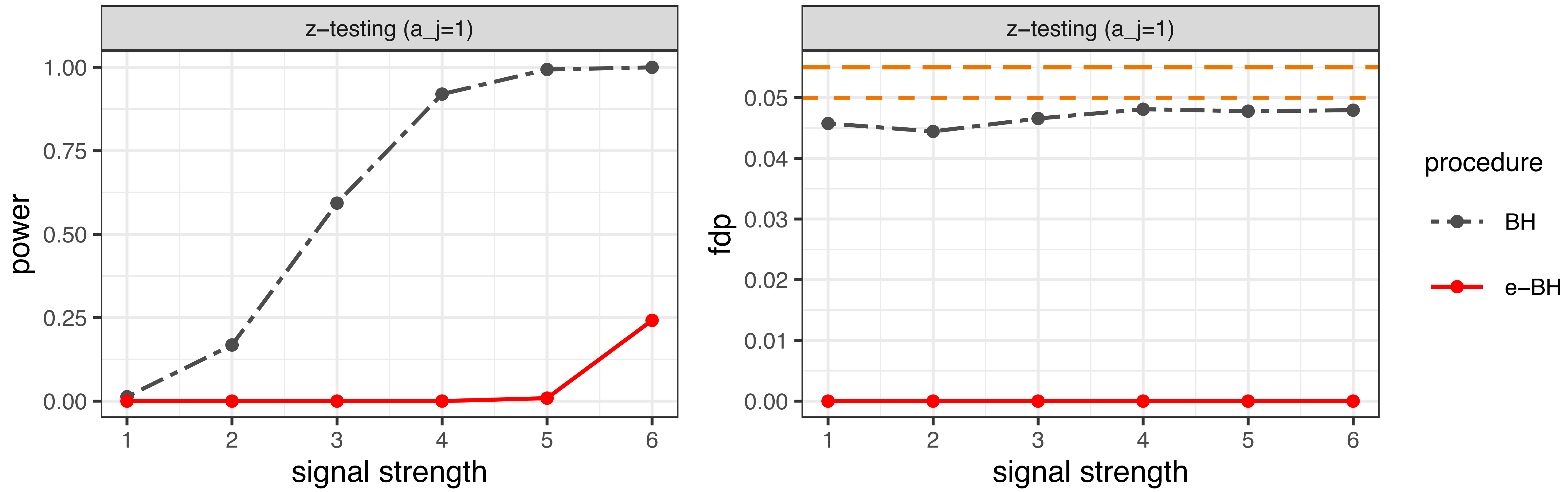
$$\leq \sum_{j \text{ null}} \mathbb{E}\left[\frac{e_j \frac{\alpha|R|}{m}}{\max(|R|,1)}\right]$$

$$\leq \frac{\alpha}{m} \sum_{j \text{ null}} \mathbb{E}[e_j]$$

$$\leq \alpha$$

All problems solved?

Actually, e-BH often exhibits lower power in practice



Testing Gaussian mean $\mu = (\mu_1, \cdots, \mu_m)$ at $\alpha = 0.05$

$H_j : \mu_j = 0$ vs. $H_j^{\text{alt}} : \mu_j > 0$;

LR e-value: $\mathrm{d}\mathcal{N}(1,1)/\mathrm{d}\mathcal{N}(0,1)$

Why the power loss?

- The e-value itself
  - only uses first-moment information
- The e-BH procedure
  - is agnostic to the joint distribution

Can we do better with partial distributional information?

Finding and filling the gap

Revisit the FDR control of e-BH

$$\mathrm{FDR} = \sum_{j \text{ null}} \mathbb{E}\left[\frac{\mathbf{1}\{j \in R\}}{\max(|R|,1)}\right] = \sum_{j \text{ null}} \mathbb{E}\left[\frac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)}\right]$$

for $t > 0$, $\mathbf{1}\{X \geq t\} \leq \frac{X}{t}$ $\longrightarrow$

$$\leq \sum_{j \text{ null}} \mathbb{E}\left[\frac{e_j \frac{\alpha|R|}{m}}{\max(|R|,1)}\right]$$

$$\leq \frac{\alpha}{m} \sum_{j \text{ null}} \mathbb{E}[e_j]$$

$$\leq \alpha$$

- This is step is tight only when $e_j \in \left\{0, \frac{m}{\alpha|R|}\right\}$
- **Our idea:** improve the power of e-BH by filling this gap

Finding and filling the gap

$$\sum_{j \text{ null}} \mathbb{E}\left[\frac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)}\right] \leq \sum_{j \text{ null}} \mathbb{E}\left[\frac{e_j \frac{\alpha|R|}{m}}{\max(|R|,1)}\right]$$

$$\text{gap per } j = \left(\frac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)} - \frac{e_j \frac{\alpha|R|}{m}}{\max(|R|,1)}\right)$$

$$\propto \left(\underline{\frac{m}{\alpha}\frac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)}} - e_j\right) \leq 0$$

key observation: $e_j' = \frac{m}{\alpha}\frac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)}$ is an e-value,

$$R(e_1', \ldots, e_m') = R(e_1, \ldots, e_m)$$

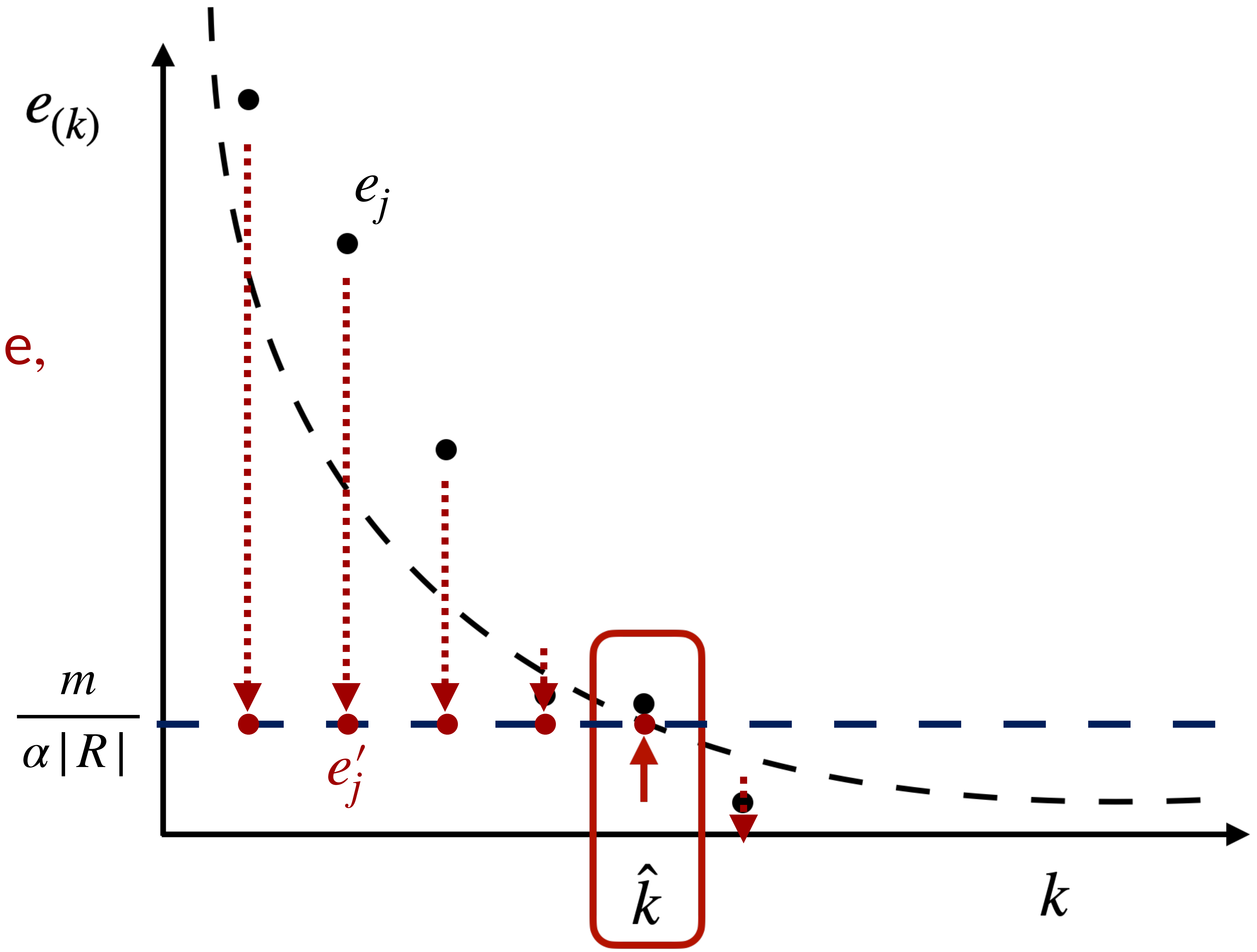
but when gap is large, $\mathbb{E}[e_j']$ is much less than $1\ldots$

- This is step is tight only when $e_j \in \left\{0, \frac{m}{\alpha|R|}\right\}$
- **Our idea:** improve the power of e-BH by filling this gap

Finding and filling the gap

key observation: $e_j' = \dfrac{m}{\alpha} \dfrac{\mathbf{1}\{e_j \geq \frac{m}{\alpha|R|}\}}{\max(|R|,1)}$ is an e-value,

$$R(e_1', \ldots, e_m') = R(e_1, \ldots, e_m)$$



Filling in the gap with conditional calibration

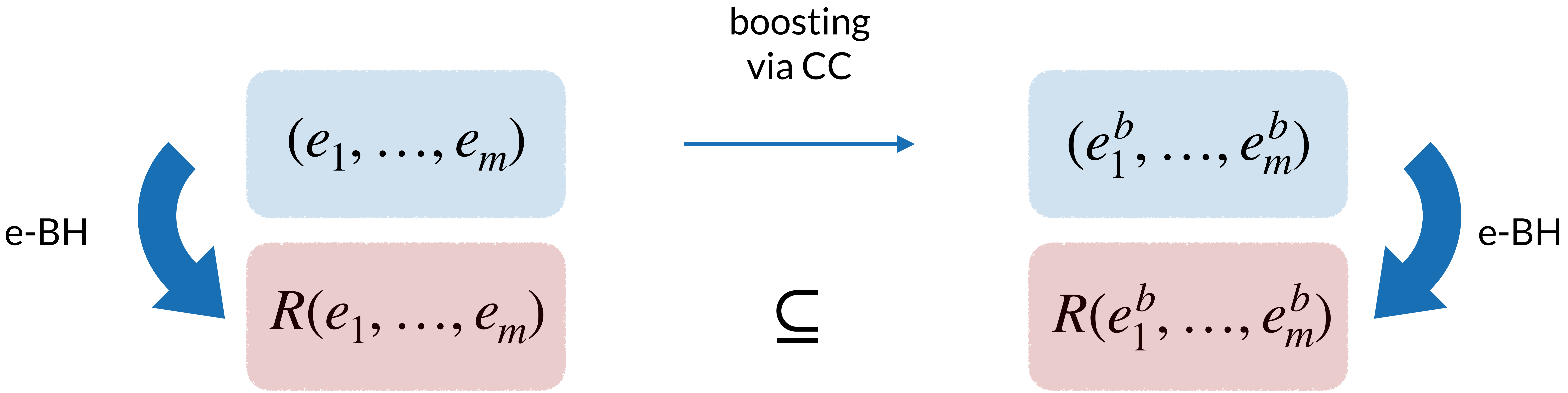
For each $j \in [m]$:

- Suppose we identify a "sufficient" statistic S_j such that we can sample from $(e_1, \ldots, e_m) \mid S_j$ under the null hypothesis H_j
- Define the quantity $\phi_j(c; S_j) = \mathbb{E}\left[\frac{m}{\alpha} \frac{\mathbf{1}\left\{ce_j \geq \frac{m}{\alpha |R \cup \{j\}|}\right\}}{|R \cup \{j\}|} - e_j \mid S_j\right]$ $\longrightarrow$ * when $c = 1$, this is the gap * increasing in c
- Find the critical value $\hat{c}_j = \sup\{c > 0 : \phi_j(c; S_j) \leq 0\}$
- Construct the boosted e-values $e_j^b = \frac{m}{\alpha} \frac{\mathbf{1}\left\{\hat{c}_j e_j \geq \frac{m}{\alpha |R \cup \{j\}|}\right\}}{|R \cup \{j\}|}$ $\longrightarrow$ * at least as big as e_j' * closes the gap to $\mathbb{E}[e_j]$

* assume for simplicity that $\phi_j(c; S_j)$ is continuous in c

* [Fithian and Lei '22] uses conditional calibration to achieve FDR control in BH

e-BH with Conditional Calibration (e-BH-CC)



e-BH-CC: filling in the gap

Validity

Theorem (L. and Ren '24). When $(e_1, \ldots, e_m)$ are e-values, the boosted e-values $(e_1^b, \ldots, e_m^b)$ are also e-values.

Power guarantee

Theorem (L. and Ren '24). Given e-values $(e_1, \ldots, e_m)$, and the boosted e-values $(e_1^b, \ldots, e_m^b)$, we have $R(e_1^b, \ldots, e_m^b) \supseteq R(e_1, \ldots, e_m)$, where each rejection set comes from running the e-BH procedure at the same level $\alpha \in (0,1)$.

e-BH-CC: computing the boost

$$\phi_j(c; S_j) = \mathbb{E}\left[\frac{m}{\alpha}\frac{\mathbf{1}\left\{ce_j \geq \frac{m}{\alpha|R\cup\{j\}|}\right\}}{|R\cup\{j\}|} - e_j \mid S_j\right]$$

$$\hat{c}_j = \sup\{c > 0 : \phi_j(c; S_j) \leq 0\}$$

- To construct $e_j^b = \frac{m}{\alpha}\frac{\boxed{\mathbf{1}\left\{\hat{c}_j e_j \geq \frac{m}{\alpha|R\cup\{j\}|}\right\}}}{|R\cup\{j\}|}$, we only need to evaluate the indicator
  - By defn. of $\hat{c}_j$: $\hat{c}_j e_j \geq \frac{m}{\alpha|R\cup\{j\}|} \iff \phi_j\left(\frac{m}{\alpha|R\cup\{j\}|}/e_j; S_j\right) \leq 0$
  - can evaluate this conditional <u>expectation</u> $\phi_j(\,\cdot\,; S_j)$ using Monte-Carlo methods
- e.g., use CIs to make approximate $e_j^b = \frac{m}{\alpha}\frac{\mathbf{1}\left\{\forall x \in \mathrm{CI}, x \leq 0\right\}}{|R\cup\{j\}|}$

(Informal) Proposition (L. and Ren '24). Running e-BH on the collection of approximated e_j^b at target FDR level α and Monte-Carlo error budget α_0, we have FDR control at $\alpha + \alpha_0$.

Using e-BH-CC: the workflow

*Identify $\{e_j\}_{j\in[m]}$

Identify S_j and $(e_1, \ldots, e_m) \mid S_j$

resampling + Monte-Carlo

Compute $\{e_j^b\}_{j\in[m]}$

Run e-BH

Rejection set $R(e_1^b, \ldots, e_m^b)$

Tailored for specific examples

Standardized

Example: testing Gaussian means with known Σ

e-values

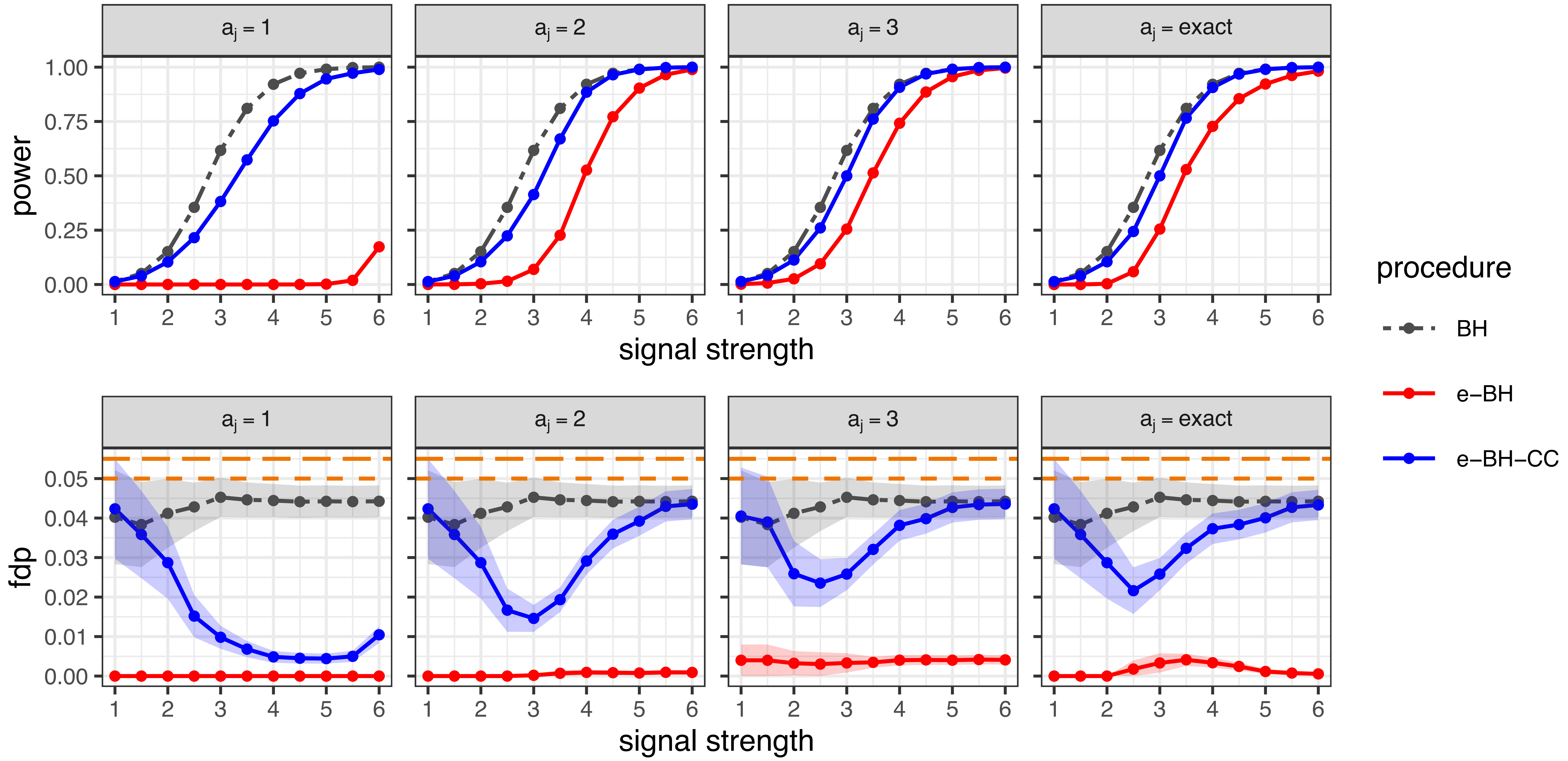
- $Z \sim \mathcal{N}_m(\mu, \Sigma)$, where Σ is known and $\Sigma_{j,j} = 1$
- $H_j: \ \mu_j = 0$ vs. $H_j^{\text{alt}}: \ \mu_j > 0$
- $e_j = \exp(a_j Z_j - a_j^2/2)$ is an e-value, for some $a_j > 0$

Have similar results for when Σ is known up to a multiplicative factor

Instantiation of e-BH-CC

- $S_j = Z_{-j} - \Sigma_{-j,j} Z_j$
- $Z_j \mid S_j \sim \mathcal{N}(0, \Sigma_{j,j})$ under H_j
- Sample $\tilde{Z}_j \sim \mathcal{N}(0, \Sigma_{j,j})$ and set $\tilde{Z}_{-j} = S_j + \Sigma_{-j,j} \tilde{Z}_j$
- $(\tilde{Z}_j, \tilde{Z}_{-j}) \sim (Z_j, Z_{-j}) \mid S_j$
- Construct e-values from $\tilde{Z}_1, \ldots, \tilde{Z}_m$

Gaussian means with known Σ



- $m = 100$, # of nonnulls = 10
- $\Sigma_{ij} = (-0.5)^{|i-j|}$
- e-value: $e_j = \exp(a_j Z_j - a_j^2/2)$

Example: conditional independence testing

- covariates $X \in \mathbb{R}^m$, response $Y \in \mathbb{R}$; $(X, Y) \sim P_{X,Y}$
- $H_j : X_j \perp\!\!\!\perp Y \mid X_{-j}$
- Model-X setting: P_X known or well approximated; no assumption on $P_{Y|X}$
  - More knowledge of the covariates
  - Abundant unsupervised data
- Find "non-null variables" X_j such that H_j false

Model-X knockoffs

[Barber and Candès '15; Candès, et al. '18]

- Generate a set of knockoff variables $\tilde{X}$ independent of Y given X and for j null:

$$(X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j}) \sim (\tilde{X}_j, X_{-j}, X_j, \tilde{X}_{-j})$$

- Construct feature importance statistics:

$$W_j = Z_j - \tilde{Z}_j$$

- $R^{kn} = \{j : W_j \geq T\}$, where

$$T = \inf\left\{ t > 0 : \frac{1 + \#\{W_j \leq -t\}}{1 \vee \#\{W_j \geq t\}} \leq \alpha \right\}$$

Theorem (Candès et al. '18).

Model-X knockoffs controls the FDR at level α: $\mathrm{FDR}[R^{\mathrm{kn}}] \leq \alpha$.

- Can be a highly variable procedure
- Power suffers in sparse settings

  - "threshold": # non-nulls $\approx 1/\alpha$

The e-BH interpretation of MX knockoffs [Ren and Barber '24]

Run MX knockoffs (level α_{kn})
to get $W_1, \ldots, W_m, T$

$$e_j = \frac{m\mathbf{1}\{W_j \geq T\}}{1 + \#\{W_j \leq -T\}}$$

- Valid (generalized) e-values:

$$\sum_{j \text{ null}} \mathbb{E}[e_j] \leq m$$

- $R^{\text{eBH}}(e_1, \ldots, e_m) = R^{\text{kn}}$

Implication

- Can be used for <u>aggregating multiple runs</u> of the random procedure by running e-BH on averaged e-values.

But accompanied with certain degrees of power loss

↓

Use e-BH-CC to reclaim the power loss

Instantiation of e-BH-CC

e-values

$$e_j = \frac{m\mathbf{1}\{W_j \geq T\}}{1 + \#\{W_j \leq -T\}}$$

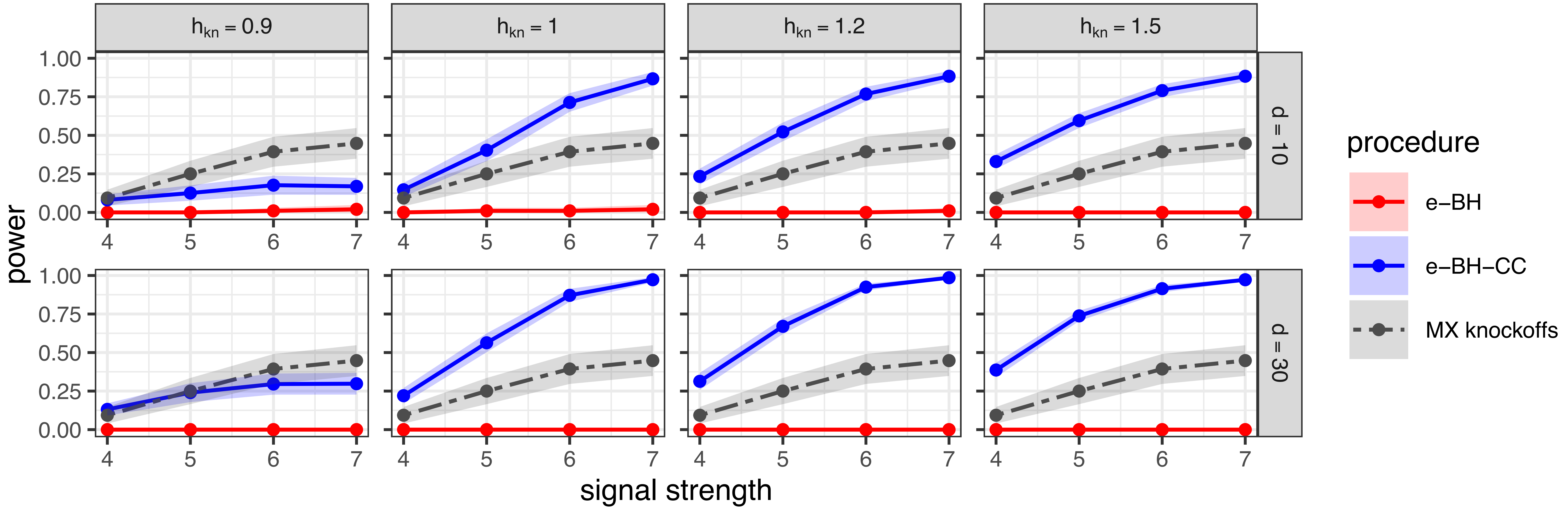
$$\bar{e}_j = \frac{1}{d}\sum_{i=1}^{d} e_j^{(i)}$$

- For each $i \in [d]$, construct $e_j^{(i)}$ from $W^{(i)}, T^{(i)}$
- averaged e-values $\Longrightarrow$ e-value

Sufficient statistics and conditional dist.

- $S_j = (X_{-j}, Y)$
- Under H_j, $X_j \mid (X_{-j}, Y)$ is simply $X_j \mid X_{-j}$
- Resample data (X', Y') from $X_j' \sim X_j \mid X_{-j}$ and setting $X'_{-j} = X_{-j}$, $Y' = Y$
- Run MX Knockoffs (d times) on resampled dataset (X', Y') and construct e-values

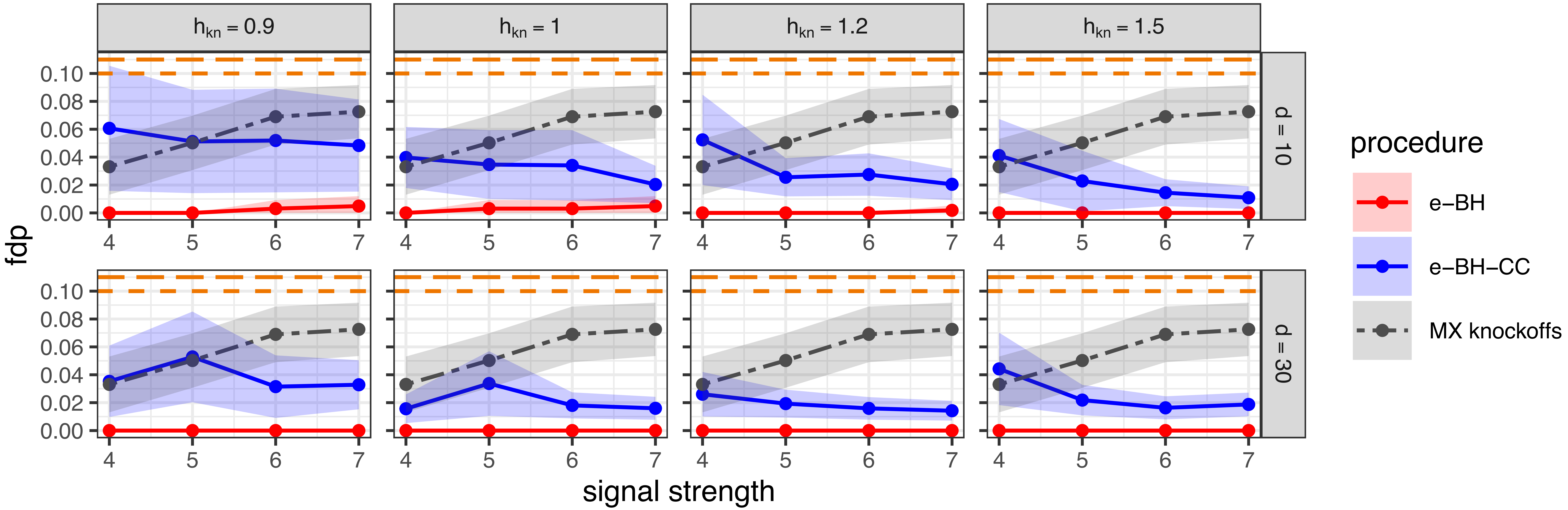
Derandomized knockoffs at the threshold - Power



Linear model: $Y = X\beta + \mathcal{N}(0,1)$

- $\alpha = 0.1$; $\alpha_{\text{kn}} = h_{\text{kn}} \cdot \alpha = 0.1 h_{\text{kn}}$
- E-values are averaged over d draws of knockoff copies
- # of non nulls = 9, $m = 200$

Derandomized knockoffs at the threshold - FDR



Linear model: $Y = X\beta + \mathcal{N}(0,1)$

- $\alpha_{\text{kn}} = h_{\text{kn}} \cdot \alpha = 0.1 h_{\text{kn}}$
- E-values are averaged over d draws of knockoff copies
- # of non nulls = 9

Example: model-free conformalized selection

- data: units $Z = (X, Y)$
- Calibration dataset: $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ (inliers)
- Test dataset: $Z_{n+1}, \ldots, Z_{n+m}$ (inliers + outliers)
- For $j \in [m]$, test $H_j : Z_{n+j} \sim Q$, where $dQ/dP = \boxed{w(X)}$ Known or identifiable

Goal: identify the outliers with FDR control

First studied by Jin and Candès '23; proposed the WCS procedure

Conformal e-values

- Assume a fixed conformity score $V(x, y)$ $\longrightarrow$ a large score suggests outlier
- $V_i = V(X_i, Y_i)$
- Conformal e-values:

$$e_j = \left(w(X_{n+j}) + \sum_{i \in [n]} w(X_i) \right) \cdot \frac{\mathbf{1}\{V_{n+j} \geq T_j\}}{w(X_{n+j}) + \sum_{i \in [n]} w(X_i) \mathbf{1}\{V_i \geq T_j\}}$$

$$T_j = \inf \left\{ t \in \{V_i\}_{i=1}^{n+m} : \frac{m}{w(X_{n+j}) + \sum_{i=1}^{n} w(X_i)} \cdot \frac{w(X_{n+j}) + \sum_{i=1}^{n} w(X_i) \mathbf{1}\{V_i \geq t\}}{\left(\sum_{k=1}^{m} \mathbf{1}\{V_{n+k} \geq t\}\right) \vee 1} \leq \alpha \right\}$$

Can make conformal p-values, but BH will not have provable FDR control! [Jin and Candès ’23]

Instantiation of e-BH-CC

- e_j are valid e-values: $\mathbb{E}[e_j] = 1$ for inliers
- Applying e-BH to $\{e_j\}_{j\in[m]}$ is (almost) equivalent to WCS [Jin and Candès '23]
  - Practically equivalent, guaranteed no less powerful

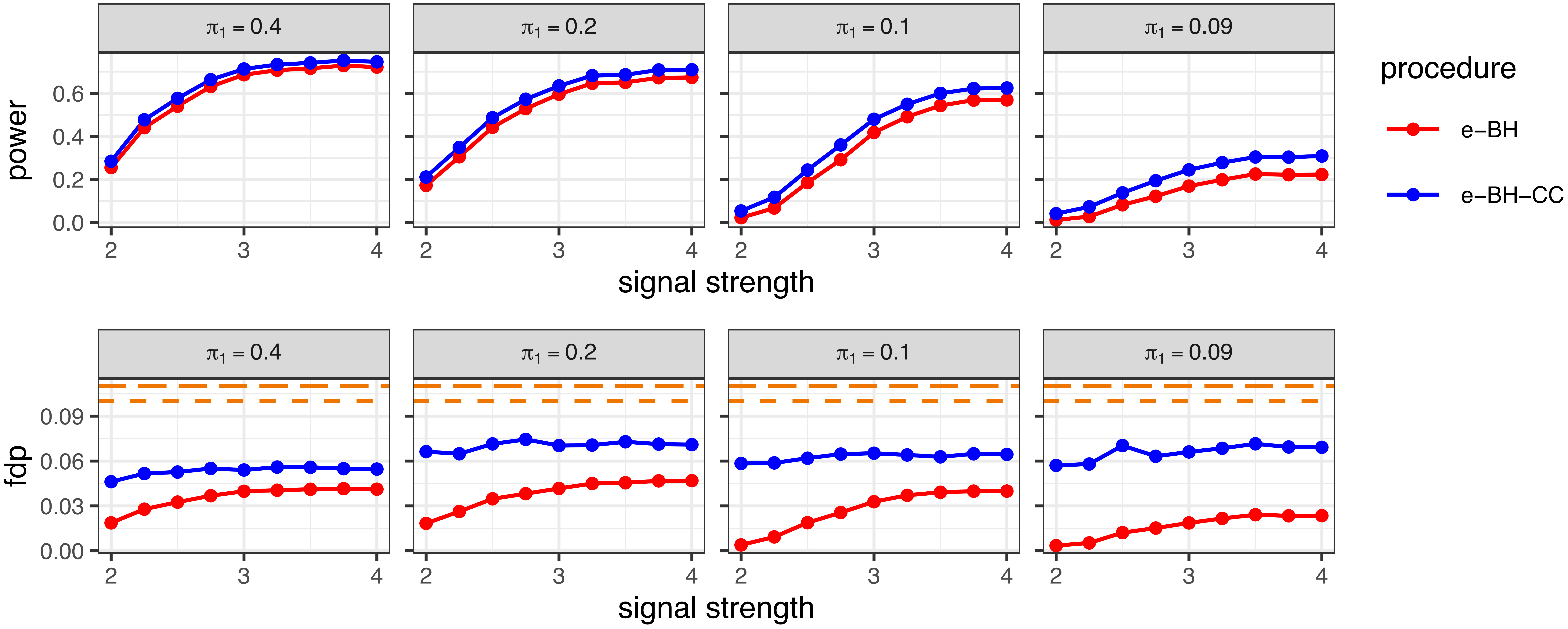
Sufficient statistics and conditional dist.

- $S_j : (\mathscr{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}})$

  un-ordered set of $\{Z_1, \ldots, Z_n\} \cup \{Z_{n+j}\}$

- $Z_{n+j} \mid \mathscr{E}_j, \{Z_{n+k}\}_{k\in[m]\setminus\{j\}} \sim \sum_{Z\in\mathscr{E}_j} \frac{w(X)}{\sum_{Z'\in\mathscr{E}_j} w(X')} \cdot \delta_Z,$

Outlier detection under covariate shift



π_1: fraction of outliers

Summary

- A framework for boosting the power of e-BH by leveraging partial distributional information
- Three concrete examples:
  - Parametric testing
  - Conditional independence testing
  - Model-free conformalized selection
- Empirically, substantial power improvement with controlled FDR

More to offer and to be done

- Application to boosting other multiple testing procedures
  - Many existing procedures have e-BH interpretation
- Beyond FDR control
- More efficient computation
  - Monte-Carlo methods

Thank you!

https://arxiv.org/abs/2404.17562