Multiple hypothesis testing with e-values and dependence

Ruodu Wang

http://sas.uwaterloo.ca/~wang

Department of Statistics and Actuarial Science
University of Waterloo

Neyman Seminar

Department of Statistics, UC Berkeley April 7, 2021 (Online)

Agenda

A little bit of what I do

A random vector $\mathbf{X} = (X_1, \dots, X_n)$

Assumptions

marginals may be known; dependence is unknown/arbitrary

Questions:

- properties of $\Psi(\mathbf{X})$ for some $\Psi : \mathbb{R}^n \to \mathbb{R}^d$
- $\mathbb{P}(\mathbf{X} \in A)$ for some $A \subseteq \mathbb{R}^n$
- "optimal" dependence structures of $\mathbf{X}$
- statistical decisions based on $\mathbf{X}$

Dates back to Fréchet-Hoeffding; has roots in Monge-Kantorovich

A little bit of what I do

Closely related problems

- Robust financial risk management
- Mass transportation
- Optimal scheduling
- Nash equilibria in resource allocation games
- Treatment effect analysis

Multiple hypothesis testing

- A (large) set of p-values is only one vector: little hope to test/verify the dependence model
- Efron'10, Large-scale Inference, p50-p51:
  "*independence among the p-values … usually an unrealistic assumption. … even PRD* [positive regression dependence] *is unlikely to hold in practice.*"
- Benjamini-Yekutieli'01: arbitrarily dependent p-values
  - Blanchard-Roquain'09, Barber-Candès'15, Fithan-Lei'20, …
- Complicated/strange dependence arises when tests statistics across experiments are generated by some adaptive procedure

Some references to e-values

Vladimir Vovk
(Royal Holloway)

Aaditya Ramdas
(Carnegie Mellon)

Bin Wang
(CAS Beijing)

- Vovk-W., E-values: Calibration, combination, and applications. arXiv:1912.06116, 2021, Annals of Statistics
- Vovk-Wang-W., Admissible ways of merging p-values under arbitrary dependence. arXiv:2007.14208, 2020
- W.-Ramdas, False discovery rate control with e-values. arXiv:2009.02824, 2020

Hypotheses testing with e-values: `http://www.alrw.net/e/`

What is an e-value?

- A hypothesis $\mathcal{H}$: a set of probability measures

Definition (e-variables and p-variables)

(1) An e-variable for testing $\mathcal{H}$ is a non-negative extended random variable $E : \Omega \to [0, \infty]$ that satisfies $\sup_{H \in \mathcal{H}} \int E \, \mathrm{d}H \le 1$.

- Realized values of e-variables are e-values.

(2) A p-variable for testing $\mathcal{H}$ is a random variable $P : \Omega \to [0, \infty)$ that satisfies $\sup_{H \in \mathcal{H}} H(P \le \alpha) \le \alpha$ for all $\alpha \in (0, 1)$.

- Realized values of p-variables are p-values.

- For simple hypothesis $\{\mathbb{P}\}$: non-negative E with mean ≤ 1
- E-test: $e(\text{data})$ large $\Longrightarrow$ reject

P-hacking

Typical scientific research

- Group A tests a medication; gets "promising but not conclusive" results
- Group B continues with new data; even more promising
- Group C continues with new data ...
- Sweep all data together to recalculate p-value ⇒ p-hacking

Many problems

- Data dependence and random stopping
- Cherry-picking
- Competitive research

What is an e-value?

- A test supermartingale: a supermartingale $X = (X_t)$ under the null with $X_0 = 1$
- Optional validity (Doob's optional stopping theorem):

$$X_\tau \text{ is an e-value for any stopping time } \tau$$

- Retrospective validity (Ville's inequality):

$$\mathbb{P}\left(\sup_{t\ge 0} X_t \ge \frac{1}{\alpha}\right) \le \alpha$$

- Bayes factors (simple hypothesis) and likelihood ratios:

$$e(\text{data}) = \frac{\Pr(\text{data} \mid \mathbb{Q})}{\Pr(\text{data} \mid \mathbb{P})}$$

- Betting scores (Shafer-Vovk'19, Shafer'21)

E for Expectation

	requirement	specific interpretation	representative forms	keyword
p-value P	$\mathbb{P}(P \le \alpha) \le \alpha$ for $\alpha \in (0,1)$	probability of a more extreme observation	$\mathbb{P}(T' \le T(\mathbf{X}) \vert \mathbf{X})$	(conditional) probability
e-value E	$\mathbb{E}^{\mathbb{P}}[E] \le 1$ and $E \ge 0$	likelihood ratios, stopped martingales, and betting scores	$\mathbb{E}^{\mathbb{P}}\left[\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \Big\vert \mathbf{X}\right]$ $\mathbb{E}^{\mathbb{P}}[M_\tau \vert \mathbf{X}]$	(conditional) expectation

An analogy of p-variables and e-variables for a simple hypothesis $\{\mathbb{P}\}$

- $\mathbf{X}$ is data
- $T(\mathbf{X})$ is any test statistic
- T' is an independent copy of $T(\mathbf{X})$ under $\mathbb{P}$
- $\mathbb{Q}$ is any probability measure
- M is a test supermartingale under $\mathbb{P}$ and τ a stopping time

(not to be confused with VanderWeele-Ding'17)

Example in testing multiple hypotheses

Multi-armed bandit problems

- K arms
- null hypothesis k: arm k has mean reward at most 1
- strategy (k_t): at time $t \geq 1$, pull arm k_t, obtain an iid reward $X_{k_t,t} \geq 0$
- aim: quickly detect arms with mean > 1
  - or maximize profit, minimize regret, etc ...
- running reward: $M_{k,t} = \prod_{j=1}^{t} X_{k,j} \mathbb{1}_{\{k_j = k\}}$
- complicated dependence due to exploration/exploitation
- $M_{1,\tau}, \ldots, M_{K,\tau}$ are e-values for any stopping time τ

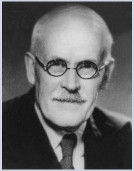
Progress

Calibration

- Admissible p-to-e calibrators
  - Power calibrators: $f_\kappa(p) = \kappa p^{\kappa-1}$ for $\kappa \in (0,1)$
  - Shafer's: $f(p) = p^{-1/2} - 1$
  - Averaging f_κ: $\int_0^1 \kappa p^{\kappa-1} \mathrm{d}\kappa = \frac{1-p+p\ln p}{p(-\ln p)^2}$
- the only admissible e-to-p calibrator: $e \to e^{-1} \wedge 1$

Sir Jeffreys

"*Users of these tests speak of the 5 per cent. point* [p-value of 5%] *in much the same way as I should speak of the* $K = 10^{-1/2}$ *point* [e-value of $10^{1/2}$], *and of the 1 per cent. point* [p-value of 1%] *as I should speak of the* $K = 10^{-1}$ *point* [e-value of 10]." (Theory of Probability, p.435, 3rd Ed.)



Calibration and combination

- The only admissible e-to-p calibrator: $e \to (1/e) \wedge 1$
- Very roughly: $p \sim 1/e$
- E-merging functions
  - arithmetic average M_K: arbitrary dependent
  - product P_K: independent, sequential
- Using $p \sim 1/e$
  - arithmetic average of e $\approx$ harmonic average of p (Wilson'19)
  - product of e $\approx$ product of p (Fisher'48)

E-merging functions

Theorem 1

Suppose that F is a symmetric e-merging function. Then $F \le \lambda + (1-\lambda)M_K$ for some $\lambda \in [0,1]$, and F is admissible if and only if $F = \lambda + (1-\lambda)M_K$ with $\lambda = F(\mathbf{0})$.

- For any symmetric e-merging function F:

$$F(\mathbf{e}) > 1 \implies M_K(\mathbf{e}) \ge F(\mathbf{e}).$$

- Asymmetric e-merging: $\mathbf{e} \mapsto \boldsymbol{\lambda} \cdot \mathbf{e}$ for $\boldsymbol{\lambda} \in \Delta_K$ where Δ_K is the standard K-simplex

---

Vovk-W., E-values: Calibration, combination, and applications. Annals of Statistics, 2021, Theorem 3.2

Connection to p-merging

Theorem 2

For any admissible p-merging function F and $\epsilon \in (0,1)$, there exist $(\lambda_1, \dots, \lambda_K) \in \Delta_K$ and admissible calibrators $f_1, \dots, f_K$ such that

$$F(\mathbf{p}) \le \epsilon \iff \sum_{k=1}^{K} \lambda_k f_k(p_k) \ge \frac{1}{\epsilon}.$$

If F is symmetric, then there exists an admissible calibrator f such that

$$F(\mathbf{p}) \le \epsilon \iff \frac{1}{K} \sum_{k=1}^{K} f(p_k) \ge \frac{1}{\epsilon}.$$

Vovk-Wang-W., Admissible ways of merging p-values under arbitrary dependence. arXiv: 2007.14208, 2020, Theorem 5.1

Merging sequential e-values

E-variables $E_1, \ldots, E_K$ are sequential if E_k is an e-variable conditional on $E_1, \ldots, E_{k-1}$ for each k.

- $\mathbb{E}[E_k \mid E_1, \ldots, E_{k-1}] \le 1$ for all $k \in \{1, \ldots, K\}$
- E-values $e_1, \ldots, e_K$ are obtained by laboratories $1, \ldots, K$
- Laboratory k makes sure that its result e_k is a valid e-value given the previous results $e_1, \ldots, e_{k-1}$

Definition (se-merging functions)

An se-merging function is an increasing Borel function $F : [0, \infty]^K \to [0, \infty]$ such that $F(E_1, \ldots, E_K)$ is an e-variable for all sequantial e-variables $E_1, \ldots, E_K$.

$$\{\text{e-merging}\} \subset \{\text{se-merging}\} \subset \{\text{ie-merging}\}$$

Test martingales

- Gaming system: a measurable function $\lambda : [0,\infty)^{<K} \to [0,1]$
- The test martingale associated with the gaming system s and initial capital $c \in [0,1]$ is the sequence $S_k : [0,\infty)^K \to [0,\infty)$ defined by $S_0 = c$ and

$$S_{k+1}(\mathbf{e}) = S_k(\mathbf{e})\big(\lambda(e_1,\dots,e_k)e_{k+1} + 1 - \lambda(e_1,\dots,e_k)\big)$$

for $k = 0,\dots,K-1$

- A martingale e-merging function is $F = S_K$ for some test martingale S.
- F and S_k are connected via

$$S_k(e_1,\dots,e_K) = F(e_1,\dots,e_k,1,\dots,1).$$

Test martingales

Theorem 3

A martingale e-merging function is an se-merging function, and each se-merging function is dominated by a martingale e-merging function (with $c = 1$).

- connection to testing via betting and confidence sequences

Test martingales

- $s = 1$ and $c = 1$: the test martingale S is given by

$$S_k(e_1, \dots, e_K) = e_1 \dots e_k,$$

and the corresponding martingale e-merging function is the product

$$F(e_1, \dots, e_K) = e_1 \dots e_K.$$

- The arithmetic mean

$$F(e_1, \dots, e_K) = \frac{e_1 + \dots + e_K}{K}$$

corresponds to the test martingale

$$S_k(e_1, \dots, e_K) = \frac{e_1 + \dots + e_k + K - k}{K}.$$

Combining sequential e-values

The general protocol

- Obtain sequential e-values $e_1, \dots, e_t, \dots$
- Decide a predictable $\lambda_1, \dots, \lambda_t, \dots \in [0,1]$
- Compute the martingale ($E_0 = 1$)

$$E_t = E_{t-1}(1 - \lambda_t + \lambda_t e_t) = \prod_{s=1}^{t}(1 - \lambda_s + \lambda_s e_s)$$

- Optimal choice of λ_t: (Waudby-Smith)-Ramdas'20
- The Kelly criterion

Progress

Testing multiple hypotheses

Basic framework

- K hypotheses $H_1, \dots, H_K$
- $\mathcal{K} = \{1, \dots, K\}$
- H_k is null if $\mathbb{P} \in H_k$
- $\mathcal{N} \subseteq \mathcal{K}$: the set of (unknown) indices of null hypotheses
- $K_0 = |\mathcal{N}|$; if $K_0/K \approx 1$ then the signals are sparse

Two settings

- H_k is associated with p-value p_k
  - p_k is realization of P_k (p-variable for $\mathbb{P}$ if $k \in \mathcal{N}$)
- H_k is associated with e-value e_k
  - e_k is realization of E_k (e-variable for $\mathbb{P}$ if $k \in \mathcal{N}$)

Testing multiple hypotheses

- A p-testing procedure $\mathcal{D} : [0,1]^K \to 2^{\mathcal{K}}$ gives the indices of rejected hypotheses based on observed p-values
- An e-testing procedure $\mathcal{D} : [0,\infty]^K \to 2^{\mathcal{K}}$ gives the indices of rejected hypotheses based on observed e-values

For a p- or e-testing procedure $\mathcal{D}$:

- $R_{\mathcal{D}}$: number of total discoveries ($R_{\mathcal{D}} = |\mathcal{D}|$)
- $F_{\mathcal{D}}$: number of false discoveries ($F_{\mathcal{D}} = |\mathcal{D} \cap \mathcal{N}|$)
- False discovery proportion (FDP): $F_{\mathcal{D}}/R_{\mathcal{D}}$ with $0/0 = 0$
- Benjamini-Hochberg'95: control the FDR $\mathbb{E}[F_{\mathcal{D}}/R_{\mathcal{D}}] \le \alpha$

$$\mathrm{FDR}_{\mathcal{D}} = \mathbb{E}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right] = \mathbb{E}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \mid R_{\mathcal{D}} > 0\right] \mathbb{P}(R_{\mathcal{D}} > 0)$$

The BH procedure

Three input ingredients:

(a) K realized p-values $p_1, \dots, p_K$ associated to $H_1, \dots, H_K$, respectively

(b) an FDR level $\alpha \in (0, 1)$

(c) (optional) dependence information or assumption on p-values, such as independence, PRDS[1] or no information

[1] PRDS: positive regression dependence on a subset, e.g., jointly Gaussian test statistics with correlations ≥ 0

The BH procedure

BH procedure

The (base) Benjamini-Hochberg (BH) procedure $\mathcal{D}(\alpha)$ rejects hypotheses with the smallest k^* p-values, where

$$k^* = \max\left\{k \in \mathcal{K} : \frac{K p_{(k)}}{k} \leq \alpha\right\}$$

with the convention $\max(\varnothing) = 0$.

	FDR	dependence
BH'95	$\frac{K_0}{K}\alpha$	independence
BY'01		PRDS
BY'01	$\ell_K \frac{K_0}{K}\alpha$	arbitrary (AD)

E-BH procedure

Three input ingredients:

(a) K realized raw e-values $e_1, \ldots, e_K$ associated to $H_1, \ldots, H_K$, respectively

(b) an FDR level $\alpha \in (0, 1)$

(c) (optional) distributional information or assumption on e-values

The e-BH procedure can be described in two steps

(1) (optional) boost the raw e-values using information in (c)

(2) apply the base e-BH procedure to the boosted e-values and level α

E-BH procedure

- $e'_1, \dots, e'_K$: raw or boosted e-values
- $e'_{[1]} \geq \dots \geq e'_{[K]}$: order statistics
- The rough relation $e \sim 1/p \Rightarrow$ use $1/e$

E-BH procedure

The base e-BH procedure $\mathcal{G}(\alpha) : [0, \infty]^K \to 2^{\mathcal{K}}$ for $\alpha > 0$ rejects hypotheses with the largest k_e^* (raw or boosted) e-values, where

$$k_e^* = \max\left\{k \in \mathcal{K} : \frac{k e'_{[k]}}{K} \geq \frac{1}{\alpha}\right\}.$$

E-BH procedure

Theorem 4

The (full) e-BH procedure has FDR at most $K_0\alpha/K$. In particular, the base e-BH procedure $\mathcal{G}(\alpha)$ directly applied to arbitrary raw e-values has FDR at most $K_0\alpha/K$.

	nice cases	general (AD)
p-BH/BY	$\frac{K_0}{K}\alpha$	penalty
e-BH	boosting	$\frac{K_0}{K}\alpha$

W.-Ramdas, False discovery rate control with e-values.
arXiv: 2009.02824, 2020, Theorem 5.1

Compliant procedures

An e-testing procedure $\mathcal{G}$ is said to be compliant at level $\alpha \in (0,1)$ if every rejected e-value e_k satisfies

$$e_k \geq \frac{K}{\alpha R_{\mathcal{G}}}.$$

- The base e-BH procedure is compliant and it dominates all other compliant procedures

Compliant procedures

Proposition 1

Any compliant e-testing procedure $\mathcal{G}$ at level α has FDR at most $\alpha K_0/K$ for arbitrary configurations of e-values.

Proof. For each $k \in \mathcal{G}$ (i.e., rejected),

$$E_k \geq \frac{K}{\alpha R_{\mathcal{G}}} \iff \frac{1}{R_{\mathcal{G}}} \leq \frac{\alpha E_k}{K}$$

The FDP of $\mathcal{G}$ satisfies

$$\frac{F_{\mathcal{G}}}{R_{\mathcal{G}}} = \frac{|\mathcal{G} \cap \mathcal{N}|}{R_{\mathcal{G}}} = \sum_{k \in \mathcal{N}} \frac{\mathbb{1}_{\{k \in \mathcal{G}\}}}{R_{\mathcal{G}}} \leq \sum_{k \in \mathcal{N}} \frac{\mathbb{1}_{\{k \in \mathcal{G}\}} \alpha E_k}{K} \leq \sum_{k \in \mathcal{N}} \frac{\alpha E_k}{K}.$$

As $\mathbb{E}[E_k] \leq 1$ for $k \in \mathcal{N}$,

$$\mathbb{E}\left[\frac{F_{\mathcal{G}}}{R_{\mathcal{G}}}\right] \leq \sum_{k \in \mathcal{N}} \mathbb{E}\left[\frac{\alpha E_k}{K}\right] \leq \frac{\alpha K_0}{K}.$$

Compliant procedures

- General compliant p-testing procedures do not have this property even if p-values are independent
- For independent p-values, a compliant p-testing procedure at α has a weaker FDR guarantee $\alpha(1+\log(1/\alpha)) > \alpha$ (Su'18)

Compliance is useful in

- data-driven structured settings
- post-selection testing
- group testing
- multi-armed bandit problems

Boosting

Define $T(x)$ as the largest value in $(K/\mathcal{K}) \cup \{0\}$ dominated by x:

$$T(x) = \frac{K}{\lceil K/x \rceil} \mathbb{1}_{\{x \geq 1\}} \text{ with } T(\infty) = K.$$

From

$$k_e^* = \max \left\{ k \in \mathcal{K} : \alpha e'_{[k]} \geq \frac{K}{k} \right\},$$

- αE_k can be safely replaced by $T(\alpha E_k)$
- It suffices to require $T(\alpha E_k)/\alpha$ to be an e-value

Boosting

For each $k \in \mathcal{K}$, take a boosting factor $b_k \geq 1$ such that

$$\max_{x \in K/\mathcal{K}} x\mathbb{P}(\alpha b_k E_k \geq x) \leq \alpha \qquad \text{if e-values are PRDS}$$

$$\mathbb{E}[T(\alpha b_k E_k)] \leq \alpha \qquad \text{otherwise (AD)}$$

and let $e'_k = b_k e_k$.

- $\mathbb{E}$ and $\mathbb{P}$ are computed under the null distribution of E_k
- Composite null: require for all probability measures in H_k
- $b_k = 1$ is always valid
- Non-linear boosting is also possible
- $\mathbf{e}'$ may not have the same order as $\mathbf{e}$.

Boosting

Example.

- For $\lambda \in (0,1)$
$$E_k = \lambda P_k^{\lambda-1},$$
where P_k is standard uniform if $k \in \mathcal{N}$
- $y_\alpha \le (\lambda^\lambda \alpha)^{1/(1-\lambda)}$
- $\lambda = 1/2 \Longrightarrow y_\alpha \le \alpha^2/2$
- $\alpha = 0.05,\ \lambda = 1/2$
  - $b_k \approx 6.32$ (AD)
  - $b_k \approx 8.94$ (PRDS)

Boosting

Example.

- For $\delta > 0$,

$$E_k = e^{\delta X_k - \delta^2/2},$$

  where X_k is standard normal if $k \in \mathcal{N}$
- $\alpha = 0.05$, $\delta = 3$
  - $b \approx 1.37$ (AD)
  - $b \approx 7.88$ (PRDS)

Progress

A multi-armed bandit problem

Problem setting

- K arms each with a reward $X^k \geq 0$
- Pulling arm k produces an iid sample $(X_1^k, X_2^k, \dots)$ from X^k
- Null hypotheses: $\mathbb{E}[X_k] \leq 1$, $k \in \mathcal{K}$
- Arms have to be pulled in order and previous arms cannot be revisited
- An arm can be pulled at most n times (budget)
- Goal: detect non-null arms as quickly as possible
- Example: investment opportunities; medical experiment

A multi-armed bandit problem

The e-value $e_{k,j}$ and the p-value $p_{k,j}$ are realized by, respectively,

$$E_{k,j} := \prod_{i=1}^{j} X_i^k \quad \text{and} \quad P_{k,j} := \left(\max_{i=1,\dots,j} E_{k,i} \right)^{-1} \quad (p \le 1/e)$$

Algorithm

- Select a p- or e-testing procedure $\mathcal{D}$ and start with $\mathbf{e} = \mathbf{p} = \mathbf{1}$
- For arm k, stop at T_k such that either $\mathcal{D}$ produces a new discovery or $T_k = n$
- Update e-values or p-values and move to arm $k+1$

The final e-variables E_k and p-variables P_k are obtained by

$$E_k = E_{k,T_k} \quad \text{and} \quad P_k = P_{k,T_k}, \quad k = 1, \dots, K.$$

A multi-armed bandit problem

Table: Conditions for the validity of the testing algorithm

	AD data across arms	AD stopping rules T_k	FDR guarantee in our experiments
e-BH	YES	YES	valid at level $\alpha K_0/K$
BH	NO	NO	not valid
BY	YES	YES	valid at level $\alpha K_0/K$
cBH	NO	YES	valid at level $\alpha K_0/K$

Consider BH, e-BH, BY and compliant BH (cBH) procedures

- BY: $\mathcal{D}(\alpha_1)$ where $\alpha_1 \ell_K = \alpha$ (Benjamini-Yekutieli'01)
- cBH: $\mathcal{D}(\alpha_2)$ where $\alpha_2(1 + \log(1/\alpha_2)) = \alpha$ (Su'18)

A multi-armed bandit problem

Data generating process

- More promising arms come first: arm k is non-null with probability $\theta(K-k+1)/(K+1)$, $\theta \in [0,1]$
- The expected number of non-nulls in this setting is $\theta/2$
- $s_k \sim \mathrm{Expo}(\mu)$ is the strength of signal for arm k
- Conditional on s_k,

$$X_1^k, \dots, X_n^k \overset{\text{iid}}{\sim} X^k = \exp\left(Z^k + s_k \mathbb{1}_{\{k \in \mathcal{K} \setminus \mathcal{N}\}} - 1/2\right)$$

  where $Z^1, \dots, Z^K$ are iid standard normal
- Set $\alpha = 0.05$ and $\theta = 0.5$ ($\Rightarrow K_0\alpha/K \approx 3.75\%$)

A multi-armed bandit problem

Table: $R = \#\{\text{rejected hypothesis}\}$, $B\% = \%(\text{unused budget})$, $\text{TD} = \#\{\text{true discoveries}\}$. Each number is computed over an average of 500 trials. Default values: $K = 500$, $n = 50$ and $\mu = 1$.

(a) Default | (b) $K = 2000$ | (c) $n = 10$

	R	$B\%$	TD	FDP%	R	$B\%$	TD	FDP%	R	$B\%$	TD	FDP%
e-BH	74.4	11.42	73.2	1.58	297.6	11.39	293.2	1.48	47.7	3.99	47.3	0.83
BH	77.0	11.44	75.3	2.13	307.8	11.41	301.4	2.07	49.3	4.01	48.7	1.06
BY	70.6	10.06	70.4	0.31	281.2	9.95	280.4	0.26	38.4	2.77	38.4	0.08
cBH	71.1	10.16	70.8	0.36	284.5	10.15	283.5	0.36	39.2	2.85	39.2	0.11

(d) $n = 100$ | (e) $\mu = 0.5$ | (f) $\mu = 2$

	R	$B\%$	TD	FDP%	R	$B\%$	TD	FDP%	R	$B\%$	TD	FDP%
e-BH	79.1	13.48	77.9	1.50	43.5	5.77	42.9	1.54	97.4	16.46	95.9	1.54
BH	81.3	13.50	79.5	2.13	46.3	5.80	45.3	2.13	99.3	16.47	97.2	2.07
BY	76.4	12.36	76.1	0.35	39.6	4.66	39.5	0.27	94.3	15.23	94.1	0.29
cBH	76.7	12.44	76.4	0.41	40.1	4.74	40.0	0.35	94.6	15.32	94.3	0.35

Correlated z-tests

- $X_k \sim \mathrm{N}(0,1)$ if $k \in \mathcal{N}$
- $X_k \sim \mathrm{N}(\delta,1)$ if $k \notin \mathcal{N}$, $\delta < 0$
- $X_1, \dots, X_K$ are jointly Gaussian
- E-values from likelihood ratios

$$E_k = \exp(\delta X_k - \delta^2/2)$$

- P-values from Neyman-Pearson tests

$$P_k = \Phi(X_k)$$

- Set $\delta = -3$

Correlated z-tests

Table: Simulation results for correlated z-tests, where $\rho_{i,j}$ is the correlation between two test statistics X_i and X_j for $i \neq j$. Each cell gives the number of rejections and, in parentheses, the realized FDP (in %). Each number is computed over an average of 1,000 trials.

(a) Independent and positively correlated tests, $K = 1000$, $K_0 = 800$

	$\rho_{ij} = 0$			$\rho_{ij} = 0.5$		
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$
BH	177.3 (8.01)	148.7 (4.07)	115.0 (1.63)	180.0 (7.00)	144.8 (3.64)	109.8 (1.50)
e-BH PRDS	171.8 (7.07)	147.6 (3.95)	114.6 (1.62)	170.2 (5.71)	142.5 (3.35)	108.0 (1.50)
BY	101.1 (1.10)	78.8 (0.57)	53.2 (0.22)	96.6 (1.03)	76.7 (0.50)	55.0 (0.20)
e-BH AD	109.4 (1.41)	85.4 (0.68)	54.6 (0.24)	103.1 (1.32)	81.4 (0.70)	56.6 (0.28)
base e-BH	97.5 (1.00)	70.6 (0.43)	36.9 (0.11)	91.9 (0.97)	69.1 (0.45)	43.6 (0.16)

Correlated z-tests

(b) Independent tests with large number of hypotheses

	$K = 20,000,\ K_0 = 10,000$			$K = 20,000,\ K_0 = 19,000$		
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$
BH	9567 (5.00)	8564 (2.49)	7164 (1.00)	681.3 (9.58)	520.2 (4.79)	357.7 (1.93)
e-BH PRDS	9092 (3.60)	8330 (2.13)	7124 (0.98)	681.3 (9.58)	509.3 (4.54)	312.1 (1.40)
BY	5956 (0.48)	4818 (0.24)	3417 (0.10)	254.1 (0.89)	177.6 (0.46)	103.1 (0.19)
e-BH AD	6811 (0.80)	5809 (0.44)	4384 (0.18)	271.0 (1.02)	159.5 (0.39)	51.4 (0.07)
base e-BH	6426 (0.64)	5234 (0.31)	3509 (0.10)	224.8 (0.69)	109.2 (0.21)	16.4 (0.01)

(c) Negatively correlated tests, $K = 1000$, $K_0 = 800$.

	$\rho_{ij} = -1/(K-1)$			$\rho_{ij} = -0.5\mathbb{1}_{\{\lvert i-j\rvert=1\}}$		
	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 2\%$
BH	177.7 (8.14)	149.0 (4.09)	115.2 (1.61)	177.2 (8.10)	148.8 (4.00)	115.3 (1.62)
e-BH PRDS	172.0 (7.13)	147.9 (3.98)	114.9 (1.59)	171.5 (7.13)	147.7 (3.89)	114.9 (1.61)
BY	101.2 (1.08)	78.8 (0.52)	53.3 (0.20)	101.3 (1.11)	78.8 (0.56)	53.2 (0.22)
e-BH AD	109.7 (1.38)	85.5 (0.65)	54.6 (0.22)	109.8 (1.40)	85.6 (0.69)	54.6 (0.24)
base e-BH	97.8 (0.98)	70.7 (0.40)	37.2 (0.11)	97.6 (0.99)	70.7 (0.41)	36.7 (0.12)

Progress

Weighted e-BH

Take $w_1, \dots, w_K \ge 0$ such that $w_1 + \cdots + w_K = K$: One can

- use $(w_1 e_1, \dots, w_K e_K)$ as the input e-values
- boost via

$$\max_{x \in K/\mathcal{K}} x\mathbb{P}(\alpha b_k E_k \ge x) \le w_k \alpha \qquad \text{if e-values are PRDS}$$

$$\mathbb{E}[T(\alpha b_k E_k)] \le w_k \alpha \qquad \text{otherwise (AD)}$$

- use random $(w_1, \dots, w_K)$ independent of the e-values with $\mathbb{E}[w_1 + \cdots + w_K] = K$ (prior information)

The same applies for compliant e-testing procedures

A class of e-testing procedures

- An increasing transform $\phi : [0,\infty] \to [0,\infty]$ is strictly increasing and continuous with $\phi(\infty) = \infty$ and $\phi(0) < 1$

E-testing procedure $\mathcal{G}(\phi)$

Define $\mathcal{G}(\phi)$ by rejecting $k^*_{e,\phi}$ hypotheses with the largest e-values, where $k^*_{e,\phi} = \max\left\{k \in \mathcal{K} : k\phi(e_{[k]})/K \geq 1\right\}$.

- $\phi : t \mapsto \alpha t \implies$ base e-BH

A class of e-testing procedures

Theorem 5

Fix $\alpha \in (0,1)$ and K. For any increasing transform ϕ, if $\mathcal{G}(\phi)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{G}(\phi)}}{R_{\mathcal{G}(\phi)}}\right] \le \alpha$$

for arbitrary configurations of e-values, then $\mathcal{G}(\phi) \subseteq \mathcal{G}(\alpha)$.

- The base e-BH procedure is optimal among $\mathcal{G}(\phi)$ with the same FDR guarantee

Applying e-BH to p-values

- A decreasing transform $\psi : [0,1] \to [0,\infty]$ is a strictly decreasing and continuous function with $\psi(0) = \infty$

P-testing procedure $\mathcal{D}(\psi)$

Define $\mathcal{D}(\psi)$ by rejecting k^*_ψ hypotheses with the largest e-values, where $k^*_\psi = \max\left\{k \in \mathcal{K} : k\psi(p_{(k)})/K \geq 1\right\}$.

- $\psi : p \mapsto \alpha/p \implies$ base BH
- equivalent to step-up methods of Benjamini-Yekutieli'01

E-BH for p-values

Proposition 2

For arbitrary p-values and a decreasing transform ψ, the testing procedure $\mathcal{D}(\psi)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi)}}{R_{\mathcal{D}(\psi)}}\right] \le \frac{K_0}{K} z_\psi,$$

where

$$z_\psi = \max_{t \in K/\mathcal{K}} t\psi^{-1}(t) \quad \textit{if p-values are PRDS},$$

$$z_\psi = \psi^{-1}(1) + \sum_{j=1}^{K-1} \frac{K}{j(j+1)} \psi^{-1}(K/j) \quad \textit{otherwise (AD)}.$$

E-BH for p-values

- For $\psi : p \to \alpha/p$,

$$\psi(p_{(k)}) \geq \frac{K}{k} \quad \Longleftrightarrow \quad \frac{Kp_{(k)}}{k} \leq \alpha.$$

- $\mathcal{D}(\psi) = \mathcal{D}(\alpha)$
- If p-values are PRDS, then $z_\psi = \alpha$ (Benjamini-Hochberg'95)
- Otherwise (Benjamini-Yekutieli'01)

$$z_\psi = \alpha + \sum_{j=1}^{K-1} \frac{\alpha}{j+1} = \alpha \ell_K$$

E-BH for p-values

(PRDS) $t \mapsto t\psi^{-1}(t)$ is decreasing on $[1, \infty) \implies z_\psi = \psi^{-1}(1)$ (D)

Proposition 3

Fix $\alpha \in (0,1)$ and K. For any decreasing transform ψ, if $\mathcal{D}(\psi)$ satisfies

$$\mathbb{E}\left[\frac{F_{\mathcal{D}(\psi)}}{R_{\mathcal{D}(\psi)}}\right] \le \alpha$$

for arbitrary configurations of PRDS p-values, then $\psi^{-1}(1) \le \alpha$. Moreover, if ψ satisfies (D), *then $\mathcal{D}(\psi) \subseteq \mathcal{D}(\alpha)$.*

- For PRDS p-values, the BH procedure is the most powerful among all $\mathcal{D}(\psi)$ satisfying (D) with the same FDR guarantee.

Progress

Some features of e-BH

The e-BH procedure

(1) works for AD e-values;

(2) requires no information on the configuration of the input e-values, and works well for weighted e-values;

(3) allows for power boosting if partial distributional information is available on some e-values;

(4) gives rise to a class of p-testing procedure which include both BH and BY as special cases;

(5) is optimal among a class of e-testing procedures under AD

Advantages of e-values

- Validity for arbitrary dependence $\Rightarrow$ expectation
- Validity for optional stopping times $\Rightarrow$ martingale
- Any p-value can be realized by sup of a continuous-time test martingale

E-values are a useful tool even if one is only interested in p-values

- Easy to combine
- Flexible to stop/continue (online testing; unfixed sample size)
- Non-asymptotic and often model-free

Ramdas-Ruf-Larsson-Koolen'20, Shafer-Shen-Vereshchagin-Vovk'11

Future work

- E-values in risk management
    - model-free e-backtesting risk measures
- FDR and other false discovery methods with p/e-values

Conjecture

Every monotone and symmetric p-testing procedure $\mathcal{D}$ with α-FDR for arbitrary dependence (like BY) is dominated by e-BH at level α applied to some calibrators.

Thank you for your attention

Working paper series on e-values: http://www.alrw.net/e/