Multiple testing

Van Hanh NGUYEN

Faculty of Information technology
Hanoi University of Agriculture

March 3, 2014

Outline

Multiple testing framework

Existing results

Perspectives

Outline

Example of single hypothesis testing

Consider the framework where one observes

- $X_1, X_2, \ldots, X_{n_1} \hookrightarrow \mathcal{N}(\mu_1, \sigma^2)$ i.i.d.
- $Y_1, Y_2, \ldots, Y_{n_2} \hookrightarrow \mathcal{N}(\mu_2, \sigma^2)$ i.i.d. and $X_i \perp Y_j$

To test hypothesis $H_0 : \{\mu_1 = \mu_2\}$ vs $H_1 : \{\mu_1 \neq \mu_2\}$, one uses a test statistic:

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- $T|H_0 \hookrightarrow T_{n_1+n_2-2}$
- Rejection region at level α: $\{|T| > t_{\frac{\alpha}{2};n_1+n_2-2}\}$
- Type I error:
  $\mathbb{P}(\text{Reject } H_0 | H_0 \text{ true}) = \mathbb{P}(|T| > t_{\frac{\alpha}{2};n_1+n_2-2} | H_0 \text{ true}) = \alpha$

Error of single hypothesis testing

	Accept H_0	**Reject** H_0
H_0 **True**	True Negative $1-\alpha$	False Positive Type I Error α
H_0 **False**	False Negative Type II Error β	True Positive $1-\beta$

$$\alpha = \mathbb{P}(\text{Type I Error}), \quad \text{Power} = 1 - \mathbb{P}(\text{Type II Error}) = 1 - \beta$$

Decision rule: among all rejection regions that have a type I error $\leq \alpha$, chose the one that has the lowest type II error.

P-value

Let F_0 be the cumulative distribution function (CDF) of test statistic T under H_0.

- One-tailed tests with rejection region $\{T > t\}$:

$$p - \text{value}(t_{obs}) = \mathbb{P}_{H_0}(T > t_{obs}) = 1 - F_0(t_{obs})$$

- One-tailed tests with rejection region $\{T < t\}$:

$$p - \text{value}(t_{obs}) = \mathbb{P}_{H_0}(T < t_{obs}) = F_0(t_{obs})$$

- Two-tailed tests with rejection region $\{|T| > |t|\}$:

$$p - \text{value}(t_{obs}) = \mathbb{P}_{H_0}(|T| > |t_{obs}|) = 1 - F_0(|t_{obs}|) + F_0(-|t_{obs}|)$$

P-value

Remark:

- If under H_0, test statistic T is a continuous variable, then p-value $P \hookrightarrow \mathcal{U}[0,1]$.
  Ex: when $P = F_0(T)$, the CDF of p-value under H_0

$$\begin{aligned} \mathbb{P}_{H_0}(P < x) &= \mathbb{P}_{H_0}(F_0(T) < x) = \mathbb{P}_{H_0}(T < F_0^{-1}(x)) \\ &= F_0(F_0^{-1}(x)) = x, \forall x \in [0,1] \end{aligned}$$

- Test at level α based on p-value: reject null hypothesis H_0 when $P < \alpha$.

P-value

Remark:

- If under H_0, test statistic T is a continuous variable, then p-value $P \hookrightarrow \mathcal{U}[0,1]$.
  Ex: when $P = F_0(T)$, the CDF of p-value under H_0

$$\begin{aligned} \mathbb{P}_{H_0}(P < x) &= \mathbb{P}_{H_0}(F_0(T) < x) = \mathbb{P}_{H_0}(T < F_0^{-1}(x)) \\ &= F_0(F_0^{-1}(x)) = x, \forall x \in [0,1] \end{aligned}$$

- Test at level α based on p-value: reject null hypothesis H_0 when $P < \alpha$.

Multiple hypothesis testing

Applications:

- Microarray analysis
- Signal detection
- Astrophysics
- ...

Example: in a microarray analysis, we aim at finding the genes that are differentially expressed between the two conditions.

- first cond.: tumor cells; second cond.: healthy cells.
- two independent samples:

$$(Y^1, \dots, Y^{n_1}) \text{ iid } \mathcal{N}(\mu_1, \Sigma) \text{ and } (Z^1, \dots, Z^{n_2}) \text{ iid } \mathcal{N}(\mu_2, \Sigma)$$

- Y_i^j (resp. Z_i^j): the expression level of the i-th gene for the j-th individual of the first (resp. second) condition.

Multiple hypothesis testing

Applications:

- Microarray analysis
- Signal detection
- Astrophysics
- ...

Example: in a microarray analysis, we aim at finding the genes that are differentially expressed between the two conditions.

- first cond.: tumor cells; second cond.: healthy cells.
- two independent samples:

$$(Y^1, \dots, Y^{n_1}) \text{ iid } \mathcal{N}(\mu_1, \Sigma) \text{ and } (Z^1, \dots, Z^{n_2}) \text{ iid } \mathcal{N}(\mu_2, \Sigma)$$

- Y_i^j (resp. Z_i^j): the expression level of the i-th gene for the j-th individual of the first (resp. second) condition.

Multiple hypothesis testing

Example:

- $\mu_1, \mu_2 \in \mathbb{R}^n$, Σ diagonal covariance matrice.
- $\{1 \leq i \leq n : \mu_{1i} \neq \mu_{2i}\} \equiv \{\text{differentially expressed genes}\}$.
- test simultaneously n hypotheses

$$H_{0,i} : "\mu_{1i} = \mu_{2i}" \text{ against } H_{1,i} : "\mu_{1i} \neq \mu_{2i}".$$

Example:

- A typical microarray experiment: test 10000 hypotheses.
- Suppose H_0 is true in all cases, use the standard p-value cut-off of 0.05 for each test.
- This procedure makes on average 500 false positives.

$\rightsquigarrow$ to correct a priori the level of the single tests in order to obtain the "quantity" of false positives that is below a nominal level α.

Multiple hypothesis testing

Example:

- $\mu_1, \mu_2 \in \mathbb{R}^n$, Σ diagonal covariance matrice.
- $\{1 \leq i \leq n : \mu_{1i} \neq \mu_{2i}\} \equiv \{\text{differentially expressed genes}\}$.
- test simultaneously n hypotheses

$$H_{0,i} : "\mu_{1i} = \mu_{2i}" \text{ against } H_{1,i} : "\mu_{1i} \neq \mu_{2i}".$$

Example:

- A typical microarray experiment: test 10000 hypotheses.
- Suppose H_0 is true in all cases, use the standard p-value cut-off of 0.05 for each test.
- This procedure makes on average 500 false positives.

$\rightsquigarrow$ to correct a priori the level of the single tests in order to obtain the "quantity" of false positives that is below a nominal level α.

"Quantity" of false positives

Table : Possible outcomes from testing n hypotheses $H^1, \ldots, H^n$.

	Accept H^i	Reject H^i	Total
H^i true	TN	FP	n_0
H^i false	FN	TP	n_1
Total	W	R	n

Approaches to control Type I Errors:

- Family-wise error rate (FWER): $\text{FWER} = \mathbb{P}(\text{FP} \geq 1)$
- False discovery rate (FDR):

$$\text{FDR} = \mathbb{E}\left[\frac{\text{FP}}{\max(\text{R},1)}\right] = \mathbb{E}\left(\frac{\text{FP}}{\text{R}}\middle|\text{R} > 0\right)\mathbb{P}(\text{R} > 0)$$

- Positive false discovery rate (pFDR): $\text{pFDR} = \mathbb{E}\left(\frac{\text{FP}}{\text{R}}\middle|\text{R} > 0\right)$

"Quantity" of false negatives

Table : Possible outcomes from testing n hypotheses $H^1, \ldots, H^n$.

	Accept H^i	Reject H^i	Total
H^i true	TN	FP	n_0
H^i false	FN	TP	n_1
Total	W	R	n

Approaches to control Type II Errors:

- False non-discovery rate (FNR):

$$\text{FNR} = \mathbb{E}\left[\frac{\text{FN}}{\max(\text{W},1)}\right] = \mathbb{E}\left(\frac{\text{FN}}{\text{W}}\middle|\text{W} > 0\right)\mathbb{P}(\text{W} > 0)$$

- Positive false non-discovery rate (pFNR):

$$\text{pFNR} = \mathbb{E}\left(\frac{\text{FN}}{\text{W}}\middle|\text{W} > 0\right)$$

Multiple testing procedure (MTP)

Definitions:

- A MTP := a random subset R of $\{1, \ldots, n\}$ that the indexes selected correspond to the rejected null hypotheses.
- The MTP based on the p-value family $p = \{p_i, 1 \leq i \leq n\}$:

$$R(p) = \{1 \leq i \leq n : p_i \leq t(p)\}.$$

- FDR control procedure: find a rejection region Γ (find a threshold $t(p)$) whose FDR $\leq \alpha$.

We shall call an FDR procedure

- valid if it controls the FDR at a level α,
- efficient (optimal) if it has the smallest FNR among all FDR procedures at level α.

Multiple testing procedure (MTP)

Definitions:

- A MTP := a random subset R of $\{1, \ldots, n\}$ that the indexes selected correspond to the rejected null hypotheses.
- The MTP based on the p-value family $p = \{p_i, 1 \leq i \leq n\}$:

$$R(p) = \{1 \leq i \leq n : p_i \leq t(p)\}.$$

- FDR control procedure: find a rejection region Γ (find a threshold $t(p)$) whose FDR $\leq \alpha$.

We shall call an FDR procedure

- valid if it controls the FDR at a level α,
- efficient (optimal) if it has the smallest FNR among all FDR procedures at level α.

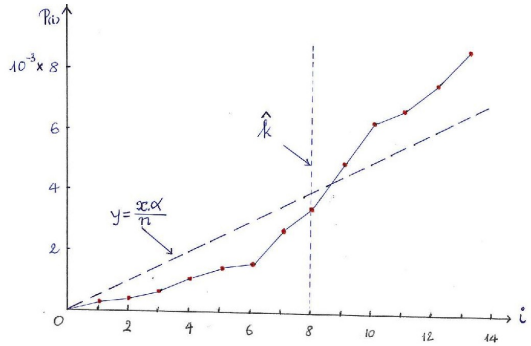
Outline

Outline

FDR control procedures

Benjamini & Hochberg (1995) procedure

- Order p-values $p_{(1)} \leq \ldots \leq p_{(n)}$.
- Let $\hat{k} = \max\{1 \leq i \leq n : p_{(i)} \leq i\alpha/n\}$
- Reject all $H^{(i)}$ for $i = 1, \ldots, \hat{k}$ (threshold $t(p) = p_{\hat{k}}$).



The BH procedure over-controls FDR: $\text{FDR}_{\text{BH}} \leq \theta\alpha$, where θ is the proportion of true null hypotheses.

$\rightsquigarrow$ apply the BH procedure at level $\alpha/\hat{\theta}$ to improve power, where $\hat{\theta}$ is an estimator of θ (Benjamini & Hochberg, 2000; Genovese & Wasserman, 2004; Blanchard & Roquain, 2009; Liang & Nettleton, 2012; ...).

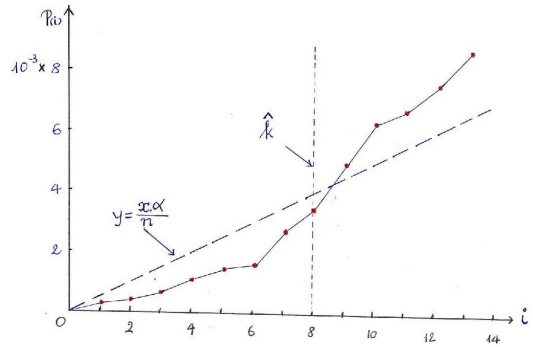
- Compute an estimator of θ as $\hat{\theta}$
- Reject all $H^{(i)}$ for $i = 1, \cdots, \hat{l}$, where

$$\hat{l} = \max\{i : p_{(i)} \leq \frac{i\alpha}{n\hat{\theta}}\} \geq \hat{k}.$$

FDR control procedures

Benjamini & Hochberg (1995) procedure

- Order p-values $p_{(1)} \leq \ldots \leq p_{(n)}$.
- Let $\hat{k} = \max\{1 \leq i \leq n : p_{(i)} \leq i\alpha/n\}$
- Reject all $H^{(i)}$ for $i = 1, \ldots, \hat{k}$ (threshold $t(p) = p_{\hat{k}}$).



The BH procedure over-controls FDR: $\text{FDR}_{\text{BH}} \leq \theta\alpha$, where θ is the proportion of true null hypotheses.

$\rightsquigarrow$ apply the BH procedure at level $\alpha/\hat{\theta}$ to improve power, where $\hat{\theta}$ is an estimator of θ (Benjamini & Hochberg, 2000; Genovese & Wasserman, 2004; Blanchard & Roquain, 2009; Liang & Nettleton, 2012; ...).

- Compute an estimator of θ as $\hat{\theta}$
- Reject all $H^{(i)}$ for $i = 1, \cdots, \hat{l}$, where

$$\hat{l} = \max\{i : p_{(i)} \leq \frac{i\alpha}{n\hat{\theta}}\} \geq \hat{k}.$$

Outline

Mixture model in multiple testing setup

Notation:

- Test simultaneously n hypotheses with p-values $P_1, \ldots, P_n$.
- Under H_0, $P_i \sim \mathcal{U}([0,1])$; under H_1, $P_i \sim F_1$ unknown
- θ: the proportion of true null hypotheses.

The CDF of P_i is a mixture:

$$F(x) = \theta x + (1-\theta)F_1(x), \text{ for } x \in [0,1],$$

and the density function of P_i is

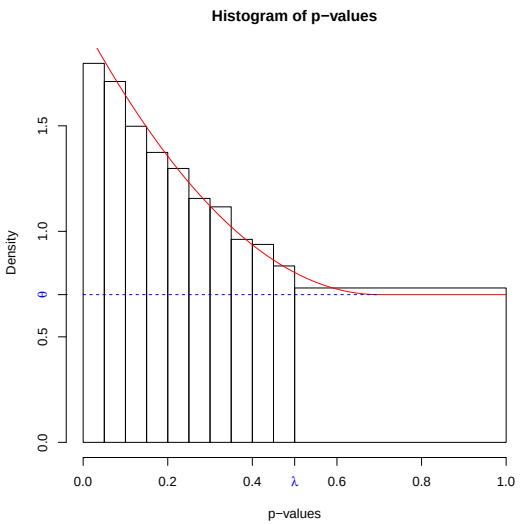
$$f(x) = \theta \mathbf{1}_{[0,1]}(x) + (1-\theta)f_1(x),$$

where f_1 is an unknown density on $[0,1]$.

- Parameters of the model: (θ, f_1).

Storey's estimators

Reasonable assumption: f_1 is non-increasing with $f_1(1) = 0$.



$\hat{f}_I$: a histogram estimator of f. Storey (2000) suggested an estimator:

$$\begin{aligned}\hat{\theta}_n(\lambda) &= \frac{\#\{P_i > \lambda : 1 \leq i \leq n\}}{n(1-\lambda)} \\ &= \hat{f}_I(x), \text{ for } x \in [\lambda, 1].\end{aligned}$$

Properties of Storey's estimator

- As λ increases, the bias decreases while the variance increases.
- $\hat{\theta}_n(\lambda)$ is unbiased if and only if $f_{1|[\lambda,1]} = 0$
- Oracle estimator: if $f_{1|[\lambda^*,1]} = 0$ then

$$\sqrt{n}\big(\hat{\theta}_n(\lambda^*) - \theta\big) \xrightarrow[n\to\infty]{d} N\big(0, \theta\big(\tfrac{1}{1-\lambda^*} - \theta\big)\big).$$

Choice of the parameter λ

- Fixed choice: use predetermined values of λ
  - $\lambda = 1/2$: the most popular choice
- Dynamic choice: use data to choose λ dynamically
  - Benjamini et al. (2000), Storey (2002), Nettleton et al. (2006); Gavrilov et al. (2009), ... choose λ dynamically
  - Celisse & Robin (2010): choose λ based on a cross-validation method

Some other estimators of θ

With the assumption $f_1(1) = 0$, estimate θ by $\hat{f}(1)$, where $\hat{f}$ is a nonparametric density estimator of f.

- Langaas et al. (2005): the Grenander estimator of monotone density
- Based on kernel density estimation, Neuvial (2013) proposes an estimator converging to θ at rate $n^{-k/(2k+1)}\eta_n$, where $\eta_n \to +\infty$ and k controls the regularity of f_1 near $x = 1$.

$\rightsquigarrow$ Can we construct an estimator of θ converging at parametric rate? with optimal asymptotic variance?

Our results

Define, for $\lambda^* \in (0, 1]$

$$\mathcal{F}_{\lambda^*} = \{f_1 : [0,1] \mapsto \mathbb{R}^+, \text{ non increasing density, positive on } [0, \lambda^*) \text{ and such that } f_{1|[\lambda^*,1]} = 0\}.$$

Two different cases: $\lambda^* < 1$ and $\lambda^* = 1$

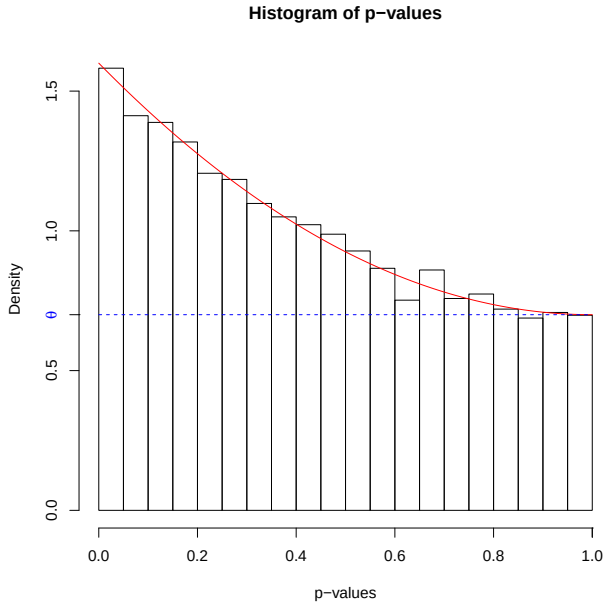
- $\lambda^* = 1$: It does not exist any estimator of θ converging at parametric rate.
- $\lambda^* < 1$: we can construct estimators converging at parametric rate but they are not asymptotically efficient (i.e. attain the optimal asymptotic variance).

Case $\lambda^* < 1$: estimators with parametric rate (I)

A histogram based estimator

$\hat{f}_I$: a histogram estimator of f.
Define an estimator of θ as

$$\hat{\theta}_{I,n} = \min_{x \in [0,1]} \hat{f}_I(x)$$



Theorem

Suppose that $f_1 \in \mathcal{F}^*_\lambda$ with $\lambda^* < 1$ and I is fine enough, then the estimator $\hat{\theta}_{I,n}$ has the following properties

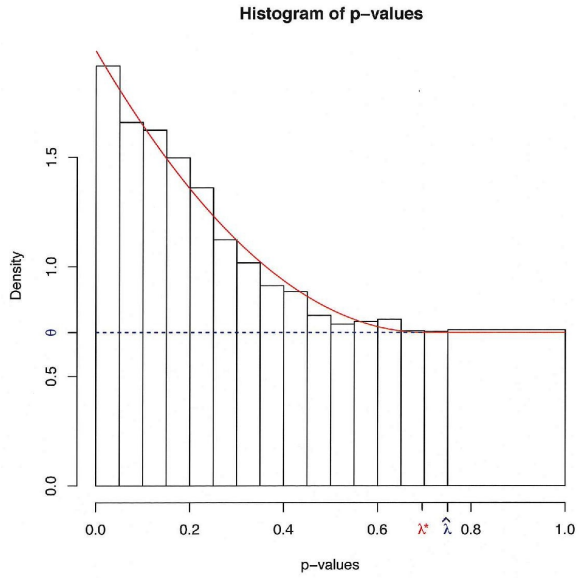
i) *$\hat{\theta}_{I,n}$ converges almost surely to θ,*

ii) $\limsup_{n \to \infty} n\mathbb{E}\left[(\hat{\theta}_{I,n} - \theta)^2\right] < +\infty.$

Case $\lambda^* < 1$: estimators with parametric rate (II)

Celisse & Robin (2010)'s procedure

$\hat{\theta}_n^{CR}$: estimator proposed by Celisse & Robin (2010)
$\hat{\lambda}$: chosen adaptively based on cross-validation method.



Theorem

Under some assumptions, the estimator $\hat{\theta}_n^{CR}$ has the following properties

i) $\hat{\theta}_n^{CR}$ *converges almost surely to* θ,

ii) $\hat{\theta}_n^{CR}$ *is* $\sqrt{n}$*-consistent, i.e.* $\sqrt{n}(\hat{\theta}_n^{CR} - \theta) = O_{\mathbb{P}}(1)$,

iii) *If the parameter* p *in leave-p-out estimator is fixed then* $\limsup_{n\to\infty} n\mathbb{E}\left[(\hat{\theta}_n^{CR} - \theta)^2\right] < +\infty$.

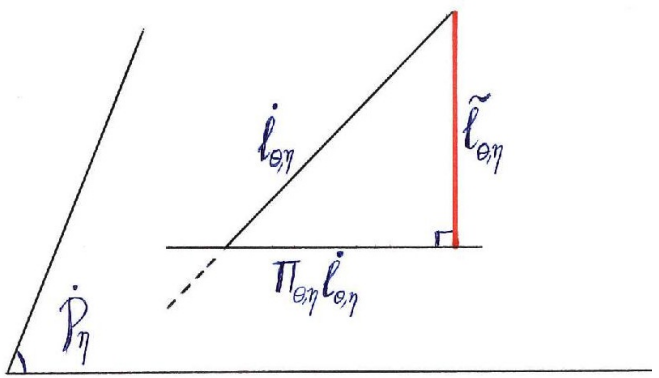
Asymptotic efficiency

In semi-parametric setup: $\mathcal{P} = \{\mathbb{P}_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{F}\}$, with $\Theta \subset \mathbb{R}$ an open set and $\mathcal{F}$ an infinite dimension set of densities.

- The ordinary score function: $\dot{l}_{\theta,\eta} = \frac{\partial}{\partial\theta} \log d\mathbb{P}_{\theta,\eta}$.
- A tangent set for η:

$$\dot{\mathcal{P}}_\eta = \left\{ \frac{\partial}{\partial t}\Big|_{t=0} \log d\mathbb{P}_{\theta,\eta_t} : \text{ for suitable paths } t \mapsto \eta_t \text{ in } \mathcal{F} \right\}$$

- The efficient score function: $\tilde{l}_{\theta,\eta} = \dot{l}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{l}_{\theta,\eta}$, where $\Pi_{\theta,\eta}$ is the orthogonal projection onto $\overline{\text{lin}}\dot{\mathcal{P}}_\eta$ in $\mathbb{L}_2(\mathbb{P}_{\theta,\eta})$.
- The efficient information: $\tilde{I}_{\theta,\eta} = \mathbb{E}_{\theta,\eta}\tilde{l}^2_{\theta,\eta}$



Asymptotic efficiency

Definition: an estimator $\hat{\theta}_n$ is asymptotically efficient (asympt. eff.) if and only if it satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_{\theta,\eta}^{-1} \tilde{l}_{\theta,\eta}(X_i) + o_{\mathbb{P}_{\theta,\eta}}(1).$$

- By the central limit theorem and Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\mathbb{P}_{\theta,\eta}}{\rightsquigarrow} N(0, \tilde{I}_{\theta,\eta}^{-1}).$$

- The LAM theorem: the optimal variance is $\tilde{I}_{\theta,\eta}^{-1}$.

Asymptotic efficiency

In our mixture model:

$$\mathcal{P}_{\lambda^*} = \Big\{\mathbb{P}_{\theta,f_1}; \frac{d\mathbb{P}_{\theta,f_1}}{d\mu} = \theta + (1-\theta)f_1; (\theta, f_1) \in (0,1) \times \mathcal{F}_{\lambda^*}\Big\}.$$

Proposition: The efficient score function $\tilde{l}_{\theta,f_1}$ and the efficient information $\tilde{I}_{\theta,f_1}$ for estimating θ in model $\mathcal{P}_{\lambda^*}$ are given by

$$\tilde{l}_{\theta,f_1}(x) = \frac{1}{\theta} - \frac{1}{\theta[1-\theta(1-\lambda^*)]}\mathbf{1}_{[0,\lambda^*)}(x) \text{ and } \tilde{I}_{\theta,f_1} = \frac{1-\lambda^*}{\theta[1-\theta(1-\lambda^*)]}.$$

Corollary

- When $\lambda^* = 1$, we have $\tilde{I}_{\theta,f_1} = 0$, then there is no estimator of θ converging at parametric rate.
- When $\lambda^* < 1$, an estimator $\hat{\theta}_n$ of θ is asympt. eff. if and only if it satisfies

$$\hat{\theta}_n = \frac{\#\{X_i > \lambda^* : 1 \le i \le n\}}{n(1-\lambda^*)} + o_{\mathbb{P}_{\theta,f_1}}(n^{-1/2}),$$

with the optimal variance equal to $\theta\big(\frac{1}{1-\lambda^*} - \theta\big)$.

Existence of asympt. eff. estimators

For an estimator $\hat{l}_{n,\theta}(\cdot) = \hat{l}_{n,\theta}(\cdot; X_1, \ldots, X_n)$ of $\tilde{l}_{\theta,f_1}$ and every sequence $\theta_n = \theta + O(n^{-1/2})$, introduce the following conditions

$$\sqrt{n}\mathbb{P}_{\theta_n,f_1}\hat{l}_{n,\theta_n} \xrightarrow[n\to\infty]{\mathbb{P}_{\theta,f_1}} 0, \tag{1}$$

$$\mathbb{P}_{\theta_n,f_1}\|\hat{l}_{n,\theta_n} - \tilde{l}_{\theta_n,f_1}\|^2 \xrightarrow[n\to\infty]{\mathbb{P}_{\theta,f_1}} 0 \tag{2}$$

Proposition

- *The existence of asympt. eff. estimators of* θ $\iff$ *the existence of estimators* $\hat{l}_{n,\theta}$ *of* $\tilde{l}_{\theta,f_1}$ *satisfying* (1) *and* (2).
- *If* $\tilde{l}_{\theta,f_1}$ *is estimated through a plug-in estimate* $\hat{\lambda}_n$ *of* λ^**, then this condition is equivalent to* $\sqrt{n}(\hat{\lambda}_n - \lambda^*) = o_{\mathbb{P}}(1)$.

Existence

- Irregular models: f_1 has a jump point at λ^*, YES
- Regular models: conjecture that NO

Outline

Perspectives

- How do we choose the finest partition in Celisse & Robin's adaptive histogram procedure? (model selection?)
- What about non iid setup? (Hidden Markov models, Sun & Cai, 2009)
  - Let $H^i = \begin{cases} 0 & \text{if the null hypothesis } i \text{ is true,} \\ 1 & \text{otherwise} \end{cases}$
  - the unobservable sequence $\{H^i\}_1^n$ is a Markov chain
  - the variables $P_i|H^i, i = 1, \ldots, n$ are independent.

Thank you for your attention!