Contrôle du *False Discovery Rate* en tests multiples: conditions suffisantes et adaptivité.

G. Blanchard[1]

[1]Weierstrass Institut Berlin, Germany

Journées Statistiques du Sud, Porquerolles 18/06/09

Plan

1. Contrôle du FDR : conditions suffisantes
2. Adaptativité à π_0
3. Relation avec le problème de la classification

Plan

1. Contrôle du FDR : conditions suffisantes
2. Adaptativité à π_0
3. Relation avec le problème de la classification

Rappel : abstraction par p-values

- $\mathcal{H}$ famille d'hypothèses à tester
- pour tout $h \in \mathcal{H}$, on suppose l'existence d'une fonction p-value $p_h : \mathcal{X} \to [0,1]$ caractérisée par :

$$\forall P \in h \qquad \mathbb{P}_{X\sim P}\left[p_h(\mathbf{X}) \leq \alpha\right] \leq \alpha\,.$$

- On considère alors les procédures de tests multiples qui sont des fonctions de la famille des p-values :

$$\text{Data } \mathbf{X} \to p\text{-values } \mathbf{p} = (p_h(\mathbf{X}))_{h\in\mathcal{H}} \to R(\mathbf{p}) \subset \mathcal{H}$$

Rappel : modèles de distribution des p-values

- si h est satisfaite par P, alors p_h a une distribution stochastiquement bornée inférieurement par une variable $U([0,1])$:

$$P \in h \Rightarrow P(p_h \leq t) \leq t$$

(c'est la définition d'une p-value)

- **($\mathbf{U_0}$)** si h est satisfaite par P, alors $p_h \sim U([0,1])$
- **($\mathbf{A_1}$)** si h n'est pas satisfaite par P, alors $p_h \sim P_1$ pour une certaine loi P_1
- **(I)** indépendance : les p-values sont indépendantes
- **(RE)** modèle "random effects", à tendance Bayésienne : soient $h_1, \ldots, h_N$ des variables de Bernoulli indépendantes de paramètre $1 - \pi_0$, les p-values sont indépendantes conditionnellement aux (h_i) avec

$$p_i \sim \begin{cases} U([0,1]) & \text{si } h_i = 0\,, \\ P_1 & \text{si } h_i = 1\,. \end{cases}$$

Les hypothèses nulles vraies forment un ensemble aléatoire donné par $\{i : h_i = 1\}$.

Le *False Discovery rate* (FDR)

- Le FWER est souvent une mesure d'erreur trop stricte en pratique.
- Critère d'erreur moins restrictif : le False Discovery Rate (FDR) de Benjamini et Hochberg (1995) :

$$\mathrm{FDR}(R,P) = \mathbb{E}_{\mathbf{X}\sim P}\left[\frac{|R(\mathbf{X}) \cap \mathcal{H}_0(P)|}{|R(\mathbf{X})|}\mathbf{1}\{|R(\mathbf{X})| > 0\}\right]$$

- Particulièrement adapté aux processus de type screening :

Grand ensemble d'objets (hypothèses candidates)

↓

Selection de candidats prometteurs (screening)
sur la base de données "limitées"

↓

Étude complémentaire plus poussée
pour les candidats sélectionnés

Le *False Discovery rate* (FDR)

- Le FWER est souvent une mesure d'erreur trop stricte en pratique.
- Critère d'erreur moins restrictif : le False Discovery Rate (FDR) de Benjamini et Hochberg (1995) :

$$\mathrm{FDR}(R,P) = \mathbb{E}_{\mathbf{X}\sim P}\left[\frac{|R(\mathbf{X}) \cap \mathcal{H}_0(P)|}{|R(\mathbf{X})|}\mathbf{1}\{|R(\mathbf{X})| > 0\}\right]$$

- Particulièrement adapté aux processus de type screening :

Grand ensemble d'objets (hypothèses candidates)

↓

Selection de candidats prometteurs (screening)
sur la base de données "limitées"

↓

Étude complémentaire plus poussée
pour les candidats sélectionnés

Le *False Discovery rate* (FDR)

- Le FWER est souvent une mesure d'erreur trop stricte en pratique.
- Critère d'erreur moins restrictif : le False Discovery Rate (FDR) de Benjamini et Hochberg (1995) :

$$\mathrm{FDR}(R,P) = \mathbb{E}_{\mathbf{X}\sim P}\left[\frac{|R(\mathbf{X}) \cap \mathcal{H}_0(P)|}{|R(\mathbf{X})|}\mathbf{1}\{|R(\mathbf{X})| > 0\}\right]$$

- Particulièrement adapté aux processus de type screening :

Grand ensemble d'objets (hypothèses candidates)

↓

Selection de candidats prometteurs (screening)
sur la base de données "limitées"

↓

Étude complémentaire plus poussée
pour les candidats sélectionnés

Self-consistance

- Supposons que l'on sache *a priori* que $|R| \geq k$.
- Alors $FDR(R) \leq k^{-1}\mathbb{E}\left[\text{Nb d'erreurs de type I de } R\right]$
- Si R est une procédure de rejet sous le seuil t, $R = \{h : p_h \leq t\}$, alors

$$\mathbb{E}\left[\text{Nb d'erreurs de } R\right] = \sum_{h \in \mathcal{H}_0} \mathbb{P}\left[p_h \leq t\right] \leq |\mathcal{H}_0| t \leq |\mathcal{H}| t$$

- Ainsi $t \leq \alpha k/|\mathcal{H}|$ donne lieu à $FDR(R) \leq \alpha$.
- Heuristique : soit une procédure R basée sur un seuil evt. aléatoire t ; la condition ci-dessus est vérifiée *a posteriori* si $t \leq \alpha |R|/|\mathcal{H}|$.
- Condition de "self-consistance" :

$$R \subset \{h : p_h \leq \alpha |R|\} \; (p.s.) \qquad \textbf{(SC)}$$

- c'est une condition purement algorithmique, càd une contrainte sur la fonction $R(\mathbf{p})$.
- condition introduite indépendamment dans [Finner et al.(2009)] et [Blanchard and Roquain(2008)].

Self-consistance

- Supposons que l'on sache *a priori* que $|R| \geq k$.
- Alors $FDR(R) \leq k^{-1}\mathbb{E}\left[\text{Nb d'erreurs de type I de } R\right]$
- Si R est une procédure de rejet sous le seuil t, $R = \{h : p_h \leq t\}$, alors

$$\mathbb{E}\left[\text{Nb d'erreurs de } R\right] = \sum_{h \in \mathcal{H}_0} \mathbb{P}\left[p_h \leq t\right] \leq |\mathcal{H}_0| t \leq |\mathcal{H}| t$$

- Ainsi $t \leq \alpha k / |\mathcal{H}|$ donne lieu à $FDR(R) \leq \alpha$.
- Heuristique : soit une procédure R basée sur un seuil evt. aléatoire t ; la condition ci-dessus est vérifiée *a posteriori* si $t \leq \alpha |R| / |\mathcal{H}|$.
- Condition de "self-consistance" :

$$R \subset \{h : p_h \leq \alpha |R|\} \; (p.s.) \qquad \textbf{(SC)}$$

- c'est une condition purement algorithmique, càd une contrainte sur la fonction $R(\mathbf{p})$.
- condition introduite indépendamment dans [Finner et al.(2009)] et [Blanchard and Roquain(2008)].

Self-consistance

- Supposons que l'on sache *a priori* que $|R| \geq k$.
- Alors $FDR(R) \leq k^{-1}\mathbb{E}\left[\text{Nb d'erreurs de type I de } R\right]$
- Si R est une procédure de rejet sous le seuil t, $R = \{h : p_h \leq t\}$, alors

$$\mathbb{E}\left[\text{Nb d'erreurs de } R\right] = \sum_{h \in \mathcal{H}_0} \mathbb{P}\left[p_h \leq t\right] \leq |\mathcal{H}_0| t \leq |\mathcal{H}| t$$

- Ainsi $t \leq \alpha k/|\mathcal{H}|$ donne lieu à $FDR(R) \leq \alpha$.
- Heuristique : soit une procédure R basée sur un seuil evt. aléatoire t ; la condition ci-dessus est vérifiée *a posteriori* si $t \leq \alpha |R|/|\mathcal{H}|$.
- Condition de "self-consistance" :

$$R \subset \{h : p_h \leq \alpha |R|\} \; (p.s.) \qquad \textbf{(SC)}$$

- c'est une condition purement algorithmique, càd une contrainte sur la fonction $R(\mathbf{p})$.
- condition introduite indépendamment dans [Finner et al.(2009)] et [Blanchard and Roquain(2008)].

Contrôle du FDR sous la condition **(SC)**

- On a :

$$\mathrm{FDR}(R) = \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|}\right] = \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{h \in R\}}{|R|}\right]$$
$$\leq \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha |R|\}}{|R|}\right]$$

- Supposons que pout tout $h \in \mathcal{H}_0$ on ait

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha |R|\}}{|R|}\right] \leq \alpha\,, \qquad \textbf{(DC)}$$

alors $\mathrm{FDR}(R) \leq \frac{|\mathcal{H}_0|}{|\mathcal{H}|} \leq \alpha\,.$

Contrôle du FDR sous la condition **(SC)**

- On a :

$$\mathrm{FDR}(R) = \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|}\right] = \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{h \in R\}}{|R|}\right]$$
$$\leq \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha |R|\}}{|R|}\right]$$

- Supposons que pout tout $h \in \mathcal{H}_0$ on ait

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha |R|\}}{|R|}\right] \leq \alpha\,, \qquad \textbf{(DC)}$$

alors $\mathrm{FDR}(R) \leq \frac{|\mathcal{H}_0|}{|\mathcal{H}|} \leq \alpha$.

Conditions suffisantes pour le contrôle du FDR

Proposition

Si la procédure de tests multiples R satisfait la condition

$$R \subset \{h : p_h \leq \alpha |R|/m\} \qquad \textbf{(SC)}$$

et que de plus on a

$$\forall h \in \mathcal{H}_0\,, c > 0\,, \qquad \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq c|R|\}}{|R|}\right] \leq c\,, \qquad \textbf{(DC)}$$

alors

$$FDR(R) \leq \frac{|\mathcal{H}_0|}{|\mathcal{H}|}\alpha \leq \alpha\,.$$

Traitement de la condition **(DC)**

Lemma

Supposons que la fonction $\mathbf{p} \mapsto |R(\mathbf{p})|$ *est décroissante. Alors, sous l'hypothèse de distribution* **(I)**, *càd indépendance des p-values, la condition (***DC***) est satisfaite :*

$$\forall h \in \mathcal{H}_0 : \qquad \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha |R|\}}{|R|}\right] \leq \alpha \,,$$

- le résultat ci-dessus est indépendant de (**SC**).
- le lemme est plus généralement vrai si les *p*-values satisfont la condition de dépendance positive "PRDS" (introduite dans Benjamini et Yekutieli 2001)

Preuve du lemme

- fixons $h \in \mathcal{H}_0$: on raisonne conditionnellement à $(p_{h'}, h' \neq h)$.
- posons $U = p_h$, on a $|R| = g(U)$ pour g décroissante.
- soit $\mathcal{U} = \{u \in [0,1] : u \leq \alpha g(u)\}$ et $u^* = \sup \mathcal{U}$, on a

$$\mathbb{E}\left[\frac{\mathbf{1}\{U \leq \alpha g(U)\}}{g(U)}\right] \leq \frac{P(U \leq u^*)}{g(u^*)} \leq \frac{u^*}{g(u^*)} \leq C\,.$$

- dans le cas PRDS, on prouve une inégalité similaire, mais où $g(U)$ est remplacé par V, une variable stochastiquement décroissante conditionnellement à U.

Preuve du lemme

- fixons $h \in \mathcal{H}_0$: on raisonne conditionnellement à $(p_{h'}, h' \neq h)$.
- posons $U = p_h$, on a $|R| = g(U)$ pour g décroissante.
- soit $\mathcal{U} = \{u \in [0,1] : u \leq \alpha g(u)\}$ et $u^* = \sup \mathcal{U}$, on a

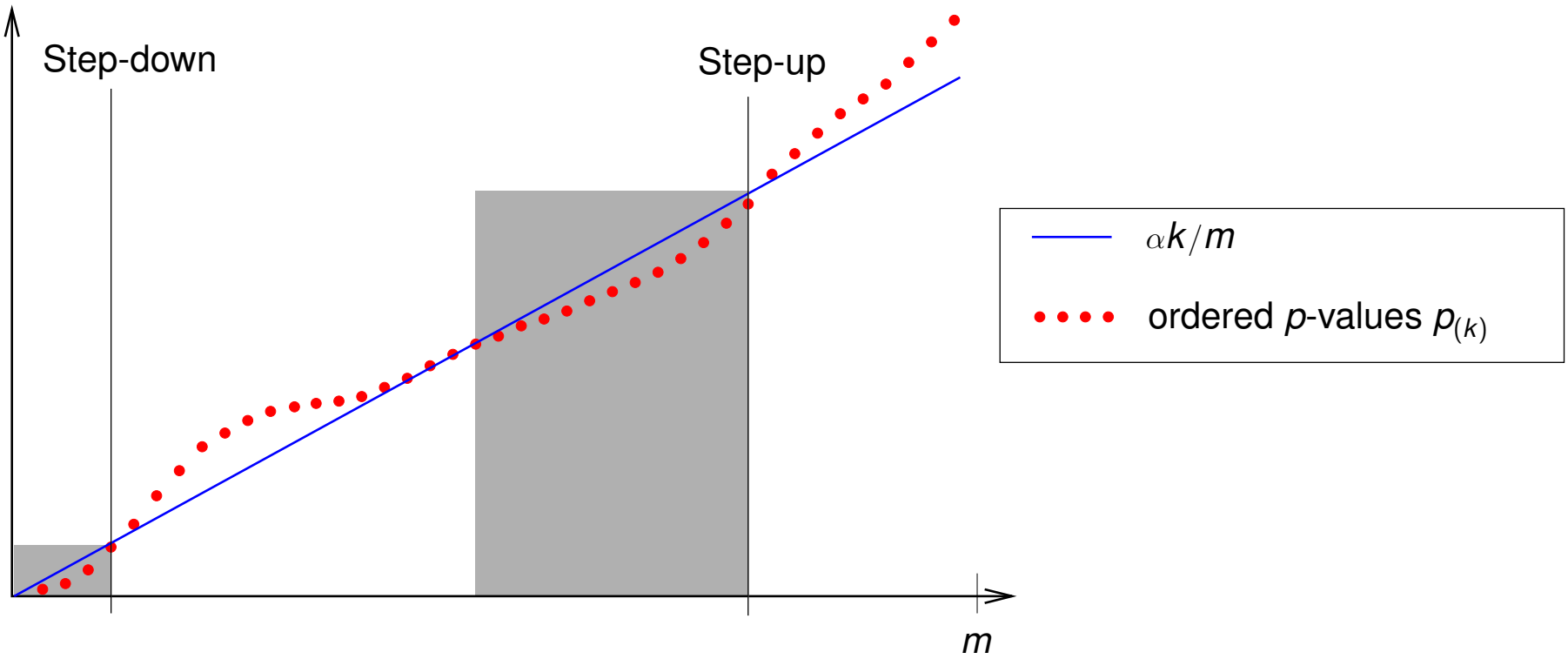
$$\mathbb{E}\left[\frac{\mathbf{1}\{U \leq \alpha g(U)\}}{g(U)}\right] \leq \frac{P(U \leq u^*)}{g(u^*)} \leq \frac{u^*}{g(u^*)} \leq C\,.$$

- dans le cas PRDS, on prouve une inégalité similaire, mais où $g(U)$ est remplacé par V, une variable stochastiquement décroissante conditionnellement à U.

La procédure du step-up linéaire

- La procédure de step-up linéaire de Benjamini et Hochberg (1995) est donnée comme suit :
  - ordonner les p-values $p^{(1)} \leq p^{(2)} \leq \ldots \leq p^{(m)}$
  - soit $\widehat{k} = \max \left\{ i : p^{(i)} \leq \alpha\beta(i) \right\}$
  - rejeter $R = \left\{ h^{(1)}, \ldots, h^{(\widehat{k})} \right\}$
- Elle statisfait la condition **(SC)** avec égalité et c'est même le plus grand ensemble de rejet satisfaisant cette condition.

Step-up linéaire (LSU)



Intermède asymptotique

- Sous l'hypothèse de distribution **(RE)**, lorsque le nombre d'hypothèses à tester tend vers ∞, par le théorème de Glivenko-Cantelli la fonction de répartition empirique des p-values tend uniformément vers

$$F_\infty(t) = \pi_0 t + (1-\pi_0)F_1(t)\,,$$

- le seuil du step-up linéaire converge p.s. vers la plus grande solution t^* de l'équation

$$F_\infty(t) = \alpha^{-1} t$$

- on vérifie bien que le FDR asymptotique est donné (si $t^* > 0$) par

$$FDR_\infty(t^*) = \frac{\pi_0 t^*}{F_\infty(t^*)} = \pi_0 \alpha$$

Généralisations

- on peut remplacer la mesure de cardinalité $|.|$ par une mesure arbitraire finie Λ sur $\mathcal{H}$. Celle-ci mesure l'"importance relative" des erreurs de type I pour différentes hypothèses.
- on peut introduire une pondération π des *p*-values : rôle similaire à celui apparaissant dans Bonferroni pondéré
- Généralisation de la condition **(SC)** dans ce cadre :

$$R \subset \{h : p_h \leq \alpha \pi(h) \Lambda(R)\} \qquad (\mathbf{SC}(\pi, \Lambda))$$

- sous cette condition et (**DC**), on a

$$\mathrm{FDR}(R) \leq \alpha \sum_{h \in \mathcal{H}_0} \Lambda(\{h\}) \pi(h) \, .$$

- Ainsi, si $\pi = \frac{d\Pi}{d\Lambda}$ est la densité d'une probabilité Π sur $\mathcal{H}$, on a $\mathrm{FDR}(R) \leq \Pi(\mathcal{H}_0)\alpha$.

Généralisations

- on peut remplacer la mesure de cardinalité $|.|$ par une mesure arbitraire finie Λ sur $\mathcal{H}$. Celle-ci mesure l'"importance relative" des erreurs de type I pour différentes hypothèses.
- on peut introduire une pondération π des p-values : rôle similaire à celui apparaissant dans Bonferroni pondéré
- Généralisation de la condition **(SC)** dans ce cadre :

$$R \subset \{h : p_h \leq \alpha \pi(h) \Lambda(R)\} \qquad (\mathbf{SC}(\pi, \Lambda))$$

- sous cette condition et (**DC**), on a

$$\mathrm{FDR}(R) \leq \alpha \sum_{h \in \mathcal{H}_0} \Lambda(\{h\}) \pi(h) \, .$$

- Ainsi, si $\pi = \frac{d\Pi}{d\Lambda}$ est la densité d'une probabilité Π sur $\mathcal{H}$, on a $\mathrm{FDR}(R) \leq \Pi(\mathcal{H}_0)\alpha$.

Généralisations

- on peut remplacer la mesure de cardinalité $|.|$ par une mesure arbitraire finie Λ sur $\mathcal{H}$. Celle-ci mesure l'"importance relative" des erreurs de type I pour différentes hypothèses.
- on peut introduire une pondération π des p-values : rôle similaire à celui apparaissant dans Bonferroni pondéré
- Généralisation de la condition **(SC)** dans ce cadre :

$$R \subset \{h : p_h \leq \alpha \pi(h) \Lambda(R)\} \qquad (\mathbf{SC}(\pi, \Lambda))$$

- sous cette condition et (**DC)**, on a

$$\mathrm{FDR}(R) \leq \alpha \sum_{h \in \mathcal{H}_0} \Lambda(\{h\}) \pi(h) \, .$$

- Ainsi, si $\pi = \frac{d\Pi}{d\Lambda}$ est la densité d'une probabilité Π sur $\mathcal{H}$, on a $\mathrm{FDR}(R) \leq \Pi(\mathcal{H}_0)\alpha$.

p-values avec dépendances arbitraires

Considérons la généralisation de la condition **(DC)** utilisant une fonction $\beta : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\forall h \in \mathcal{H}_0, \forall \alpha > 0 : \qquad \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq c\beta(|R|)\}}{|R|}\right] \leq c\,, \qquad \textbf{(DC}(\beta)\textbf{)}$$

et la modification correspondante de la condition **(SC)** :

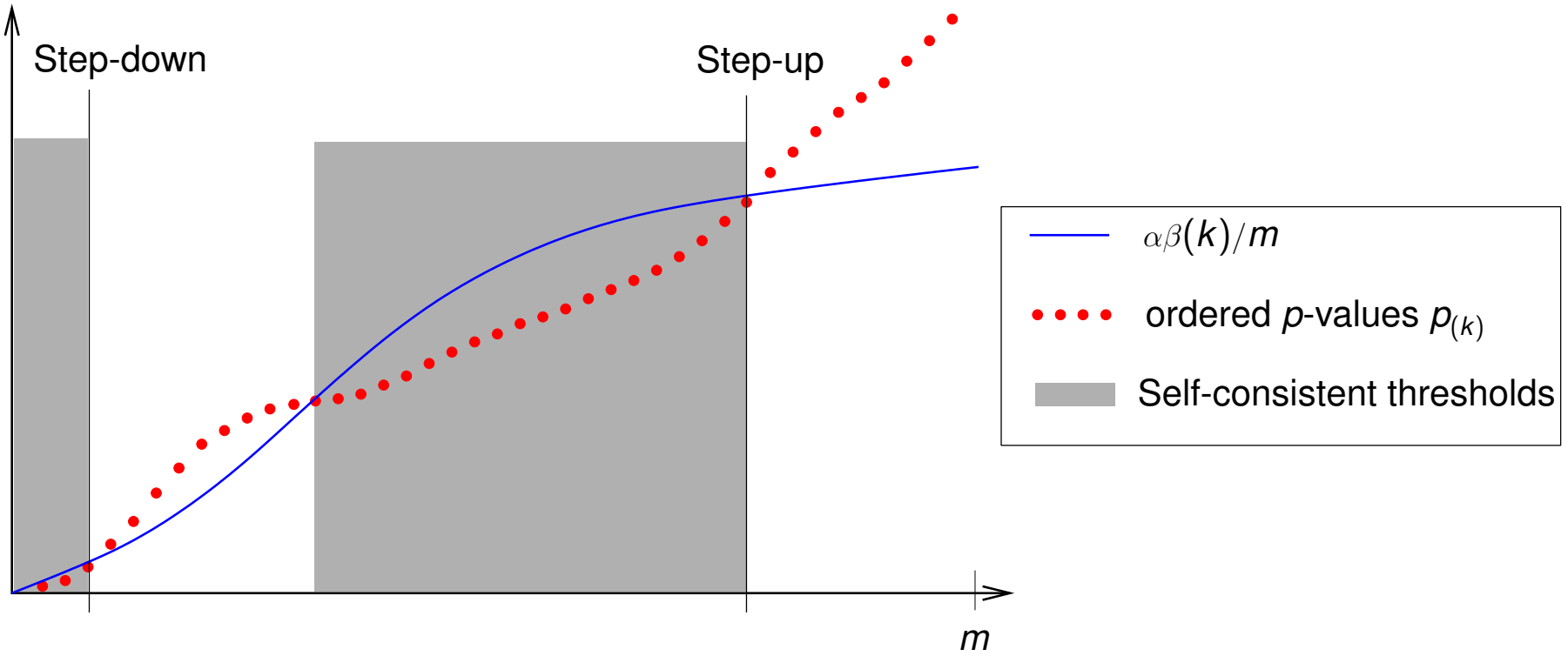
$$R \subset \{h : p_h \leq \alpha\pi(h)\beta(|R|)\} \qquad \textbf{(SC}(\alpha, \pi, \beta)\textbf{)}$$

Proposition

Sous les conditions **(SC**(α, π, β)**)** *et* **DC**(β) *on a*

$$FDR(R) \leq \alpha \sum_{h \in \mathcal{H}_0} \Lambda(\{h\})\pi(h)\,.$$

Procédures self-consistantes avec fonction de rejet β



(Condition **DC**) avec dépendances arbitraires

Lemma

Supposons que la fonction β est de la forme

$$\beta(x) = \int_0^x u d\nu(u)\,, \qquad (*)$$

où ν est une mesure de probabilité sur $\mathbb{R}_+$.
Alors pour toute procédure R , on a

$$\forall h \in \mathcal{H}_0\,, c > 0 : \qquad \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq c\beta(|R|)\}}{|R|}\right] \leq c\,,$$

Dépendances arbitraires : une famille de fonctions de rejet

- Dans le cas de dépendances arbitraires, le prix à payer est que la fonction de rejet β est toujours plus petite que la fonction linéaire du step-up linéaire.

$$\beta(x) = \int_0^x u d\nu(u)$$

- On peut construire ds contre-exemples montrant que la borne basée sur une fonction β arbitraire peut être atteinte.
- ν jour le rôle d'une sorte de "prior" sur la taille de l'ensemble d'hypothèses rejetées $|R|$.
- Le choix $\nu(i) = c^{-1} i^{-1}$ for $i = 1, \ldots, |\mathcal{H}|$ donne une autre procédure de step-up linéaire, celle de Benjamini and Yekutieli (2001). La pente est plus petite d'un facteur $c \simeq \ln |\mathcal{H}|$ par rapport au step-up linéaire "classique".

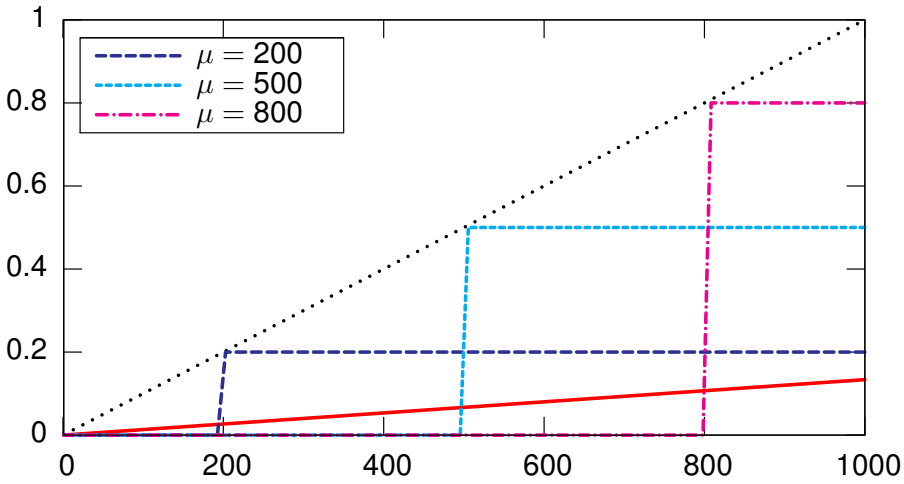
Dépendances arbitraires : une famille de fonctions de rejet

- Dans le cas de dépendances arbitraires, le prix à payer est que la fonction de rejet β est toujours plus petite que la fonction linéaire du step-up linéaire.

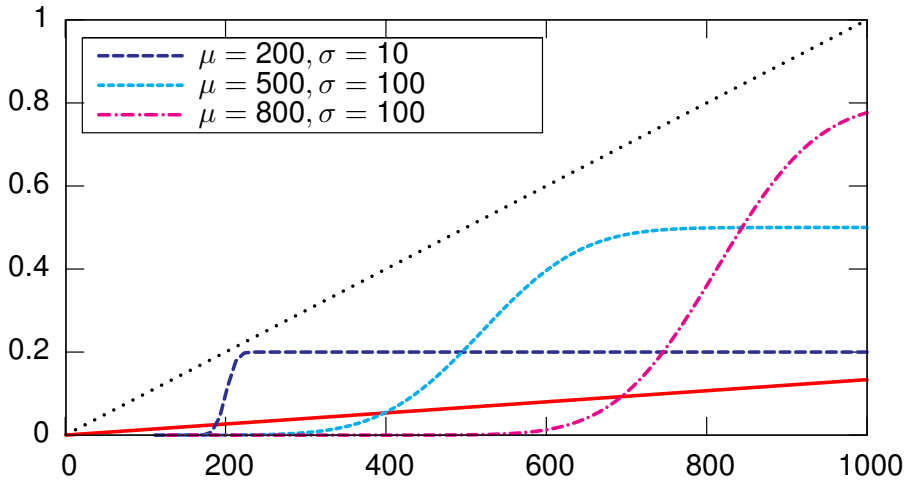
$$\beta(x) = \int_0^x u d\nu(u)$$

- On peut construire ds contre-exemples montrant que la borne basée sur une fonction β arbitraire peut être atteinte.
- ν jour le rôle d'une sorte de "prior" sur la taille de l'ensemble d'hypothèses rejetées $|R|$.
- Le choix $\nu(i) = c^{-1} i^{-1}$ for $i = 1, \ldots, |\mathcal{H}|$ donne une autre procédure de step-up linéaire, celle de Benjamini and Yekutieli (2001). La pente est plus petite d'un facteur $c \simeq \ln |\mathcal{H}|$ par rapport au step-up linéaire "classique".

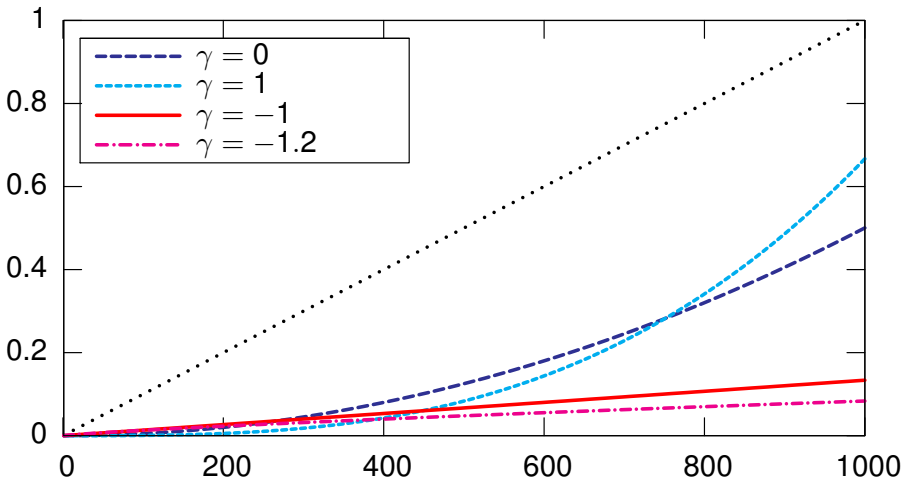
Dirac



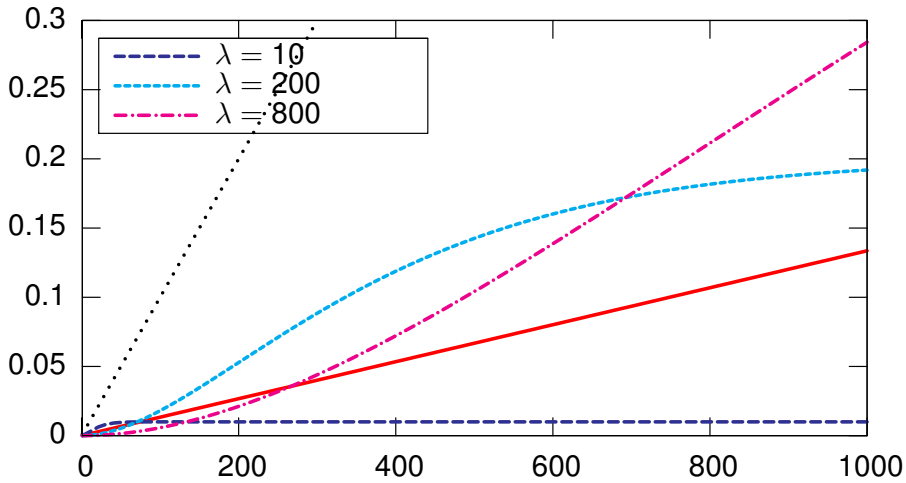
Gaussian



Power function



Exponential function



Utilité de la condition **SC** ?

- le plus grand ensemble possible de rejets sous la contrainte **(SC)** est donnée par la procédure de step-up.
- la condition “plus générale” **(SC)** a-t-elle un intérêt ?
- d’autres contraintes, par ex. géométriques, peuvent être présentes. (Ex : convexité, connexité, régularité de la frontière en imagerie)
- le step-up peut ne pas satisfaire les contraintes additionnelles, et **(SC)** offre alors plus de flexibilité.

Utilité de la condition **SC** ?

- le plus grand ensemble possible de rejets sous la contrainte **(SC)** est donnée par la procédure de step-up.
- la condition “plus générale” **(SC)** a-t-elle un intérêt ?
- d’autres contraintes, par ex. géométriques, peuvent être présentes. (Ex : convexité, connexité, régularité de la frontière en imagerie)
- le step-up peut ne pas satisfaire les contraintes additionnelles, et **(SC)** offre alors plus de flexibilité.

Ensemble continu d'hypothèses

- le cadre mathématique employé s'applique tout aussi bien à un ensemble continu d'hypothèses (par ex. processus stochastique)
- il n'est plus pertinent de supposer les p-values indépendantes
- le cadre des dépendances arbitraires ou positives-PRDS se généralise par contre sans problème. Exemple : processus Gaussien avec un opérateur de covariance positif
- La procédure de "step-up" est alors définie par

$$R = \sup \{ R \subset \mathcal{H} \ : \ R \subset \{ h : p_h \leq \alpha \pi(h) \beta(|R|) \} \} \ .$$

Plan

1. Contrôle du FDR : conditions suffisantes
2. Adaptativité à π_0
3. Relation avec le problème de la classification

- Dans les cas traités précédemment les procédures step-up considérées satisfont

$$FDR(R) \leq \pi_0 \alpha$$

où $\pi_0 = \frac{|\mathcal{H}_0|}{|\mathcal{H}|}$.

- Idéalement on voudrait remplacer la fonction de réjection β par l'"oracle"

$$\beta^* = \pi_0^{-1} \beta \ \ldots$$

- la procédure la plus intuitive consiste à utiliser un estimateur $\widehat{\pi_0}$ de π_0 comme "plug-in".

- Dans les cas traités précédemment les procédures step-up considérées satisfont

$$FDR(R) \leq \pi_0 \alpha$$

où $\pi_0 = \frac{|\mathcal{H}_0|}{|\mathcal{H}|}$.

- Idéalement on voudrait remplacer la fonction de réjection β par l'"oracle"

$$\beta^* = \pi_0^{-1} \beta \ \ldots$$

- la procédure la plus intuitive consiste à utiliser un estimateur $\widehat{\pi_0}$ de π_0 comme "plug-in".

Une procédure en 2 étapes simples (dépendances arbitraires)

- supposons que $\widehat{\pi}_0$ est une borne inférieure de confiance au niveau $1-\delta$ pour π_0
- par example, $\widehat{\pi}_0$ peut être le nombre d'hypothèses non rejetées par une première étape de test multiple avec FWER contrôlé au niveau δ .

Theorem

Si $P(\widehat{\pi}_0 > \pi) \leq \delta$ *, une procédure* R *satisfaisant la condition* **(SC)** *au niveau modifié* $\alpha\widehat{\pi}_0^{-1}$ *vérifie*

$$FDR(R) \leq \alpha + \delta\,.$$

Preuve : considérer la procédure $\widetilde{R}$ telle que

$$\widetilde{R} = R \quad \text{if} \quad \widehat{\pi}_0 < \pi_0; \qquad \widetilde{R} = \emptyset \text{ otherwise.}$$

Alors $\widetilde{R}$ vérifie la condition **SC** (au niveau $\pi_0^{-1}\alpha$) et $R = \widetilde{R}$ avec probabilité $\geq 1-\delta$.

Une procédure en 2 étapes simples (dépendances arbitraires)

- supposons que $\widehat{\pi}_0$ est une borne inférieure de confiance au niveau $1-\delta$ pour π_0
- par example, $\widehat{\pi}_0$ peut être le nombre d'hypothèses non rejetées par une première étape de test multiple avec FWER contrôlé au niveau δ .

Theorem

Si $P(\widehat{\pi}_0 > \pi) \leq \delta$ *, une procédure* R *satisfaisant la condition* **(SC)** *au niveau modifié* $\alpha\widehat{\pi}_0^{-1}$ *vérifie*

$$FDR(R) \leq \alpha + \delta\,.$$

Preuve : considérer la procédure $\widetilde{R}$ telle que

$$\widetilde{R} = R \quad \text{if} \quad \widehat{\pi}_0 < \pi_0; \qquad \widetilde{R} = \emptyset \text{ otherwise.}$$

Alors $\widetilde{R}$ vérifie la condition **SC** (au niveau $\pi_0^{-1}\alpha$) et $R = \widetilde{R}$ avec probabilité $\geq 1-\delta$.

Sous indépendance : une approche en 2 étapes

Cadre : $\widehat{\pi}_0$ est un estimateur de π_0 ; appliquer le step-up linéaire en utilisant les seuils de réjection $\alpha_i = \widehat{\pi_0}^{-1} \frac{i}{m}$.

Theorem (Essentiellement de [Benjamini et al.(2006)])

On suppose que $\widehat{\pi}_0^{-1} = G(\mathbf{p})$ est une fonction décroissante des p-values. Soit R une procédure vérifiant la modification suivante de **SC** *:*

$$R \subset \{h : p_h \leq \alpha G(\mathbf{p})|R|/m\} \ .$$

Alors, sous l'hypothèse de distribution **(I)** *(p-values indépendantes), on a :*

$$FDR(R) \leq \frac{\alpha}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[G(\mathbf{p}_{h,0})\right] \ ,$$

où $\mathbf{p}_{h,0}$ est le vecteur des p-values $\mathbf{p}$ où p_h est remplacé par 0.

Un résultat similaire est présent dans [Finner et al.(2009)].

Sous indépendance : une approche en 2 étapes

Cadre : $\widehat{\pi}_0$ est un estimateur de π_0 ; appliquer le step-up linéaire en utilisant les seuils de réjection $\alpha_i = \widehat{\pi_0}^{-1} \frac{i}{m}$.

Theorem (Essentiellement de [Benjamini et al.(2006)])

On suppose que $\widehat{\pi}_0^{-1} = G(\mathbf{p})$ *est une fonction décroissante des p-values. Soit R une procédure vérifiant la modification suivante de* **SC** *:*

$$R \subset \{h : p_h \leq \alpha G(\mathbf{p}) |R|/m\} \ .$$

Alors, sous l'hypothèse de distribution **(I)** *(p-values indépendantes), on a :*

$$FDR(R) \leq \frac{\alpha}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[G(\mathbf{p}_{h,0})\right] \ ,$$

où $\mathbf{p}_{h,0}$ *est le vecteur des p-values* $\mathbf{p}$ *où* p_h *est remplacé par 0.*

Un résultat similaire est présent dans [Finner et al.(2009)].

Estimateur arbitraire : contrôle par calibration

- soit $\widehat{\pi}_0(\mathbf{p})$ un estimateur de π_0 croissant en les p-values.
- noter (toujours sous **(I)**) que la borne $\mathbb{E}\left[\widehat{\pi}_0(\mathbf{p}_{h,0})^{-1}\right]$ est alors maximisée pour la distribution suivante :
  - les p-values des hypothèses non-nulles sont indentiquement nulles
  - les hypothèses nulles sont $U([0,1])$ (sauf une, identiquement nulle)

  Notons $DU(n,k)$ une telle distribution (avec k p-values nulles et $n-k$ uniformes)
- alors

$$\mathrm{FDR}(R) \leq \alpha \sup_{1 \leq n0 \leq n} \left(\frac{n_0}{n} \mathbb{E}_{DU(n,n-n_0+1)} \left[G(\mathbf{p}_{h,0}) \right] \right) ,$$

  qu'on peut (en principe) estimer par simulation.

L'estimateur de Storey

- l'estimateur dit de Storey (2002) (voir aussi Schweder et Spjøtvoll 1982) de π_0 , dependant du paramère $\lambda \in (0,1)$, est défini par :

$$\widehat{\pi}_0(\mathbf{p}, \lambda) = \frac{|\{h : p_h > \lambda\}|}{(1-\lambda)m} .$$

$$\widehat{\pi}_0(\mathbf{p}, \lambda) = \frac{|\{h : p_h > \lambda\}| + 1}{(1-\lambda)m} .$$

Lemma

Sous l'hypothèse de distribution **(I)** *(p-values indépendantes) :*

$$\mathbb{E}\left[\frac{(1-\lambda)m}{|\mathbf{1}\{h : p_h > \lambda\}| + 1}\right] \leq \pi_0^{-1} .$$

- Corollaire : la procédure adaptive en 2 étapes utilisant l'estimateur de Storey modifié a un FDR borné par α (sous **(I)**).

L'estimateur de Storey

- l'estimateur dit de Storey (2002) (voir aussi Schweder et Spjøtvoll 1982) de π_0 , dependant du paramère $\lambda \in (0,1)$, est défini par :

$$\widehat{\pi}_0(\mathbf{p},\lambda) = \frac{|\{h : p_h > \lambda\}|}{(1-\lambda)m} .$$

$$\widehat{\pi}_0(\mathbf{p},\lambda) = \frac{|\{h : p_h > \lambda\}| + 1}{(1-\lambda)m} .$$

Lemma

Sous l'hypothèse de distribution **(I)** *(p-values indépendantes) :*

$$\mathbb{E}\left[\frac{(1-\lambda)m}{|\mathbf{1}\{h : p_h > \lambda\}| + 1}\right] \leq \pi_0^{-1} .$$

- Corollaire : la procédure adaptive en 2 étapes utilisant l'estimateur de Storey modifié a un FDR borné par α (sous **(I)**).

L'estimateur de Storey

- l'estimateur dit de Storey (2002) (voir aussi Schweder et Spjøtvoll 1982) de π_0 , dependant du paramère $\lambda \in (0,1)$, est défini par :

$$\widehat{\pi}_0(\mathbf{p},\lambda) = \frac{|\{h : p_h > \lambda\}|}{(1-\lambda)m} .$$

$$\widehat{\pi}_0(\mathbf{p},\lambda) = \frac{|\{h : p_h > \lambda\}| + 1}{(1-\lambda)m} .$$

Lemma

Sous l'hypothèse de distribution **(I)** *(p-values indépendantes) :*

$$\mathbb{E}\left[\frac{(1-\lambda)m}{|\mathbf{1}\{h : p_h > \lambda\}| + 1}\right] \leq \pi_0^{-1} .$$

- Corollaire : la procédure adaptive en 2 étapes utilisant l'estimateur de Storey modifié a un FDR borné par α (sous **(I)**).

L'estimateur de Storey

- l'estimateur dit de Storey (2002) (voir aussi Schweder et Spjøtvoll 1982) de π_0 , dependant du paramère $\lambda \in (0,1)$, est défini par :

$$\widehat{\pi}_0(\mathbf{p},\lambda) = \frac{|\{h : p_h > \lambda\}|}{(1-\lambda)m} .$$

$$\widehat{\pi}_0(\mathbf{p},\lambda) = \frac{|\{h : p_h > \lambda\}| + 1}{(1-\lambda)m} .$$

Lemma

Sous l'hypothèse de distribution **(I)** *(p-values indépendantes) :*

$$\mathbb{E}\left[\frac{(1-\lambda)m}{|\mathbf{1}\{h : p_h > \lambda\}| + 1}\right] \leq \pi_0^{-1} .$$

- Corollaire : la procédure adaptive en 2 étapes utilisant l'estimateur de Storey modifié a un FDR borné par α (sous **(I)**).

La courbe de réjection asymptotiquement optimale de Finner et al.

- considérons l'estimateur de Storey avec un paramètre λ dépendant des données.
- on veut choisir $\lambda = p^{(i)}$, où $p^{(i)}$ sera précisément la plus grande p-value rejetée.
- Ceci donne

$$\widehat{\pi}_0(\mathbf{p}, p^{(i)}) = \frac{m-i}{(1-p^{(i)})m} .$$

- et suggère la courbe de réjection donnée par les seuils

$$\alpha_i = \alpha \frac{i}{m-(1-\alpha)i} .$$

- problème : comme $\alpha_m = 1$, la procédure step-up utilisant ces seuils rejette systématiquement toutes les hypothèses.

La courbe de réjection asymptotiquement optimale de Finner et al.

- considérons l'estimateur de Storey avec un paramètre λ dépendant des données.
- on veut choisir $\lambda = p^{(i)}$, où $p^{(i)}$ sera précisément la plus grande p-value rejetée.
- Ceci donne

$$\widehat{\pi}_0(\mathbf{p}, p^{(i)}) = \frac{m - i}{(1 - p^{(i)})m} .$$

- et suggère la courbe de réjection donnée par les seuils

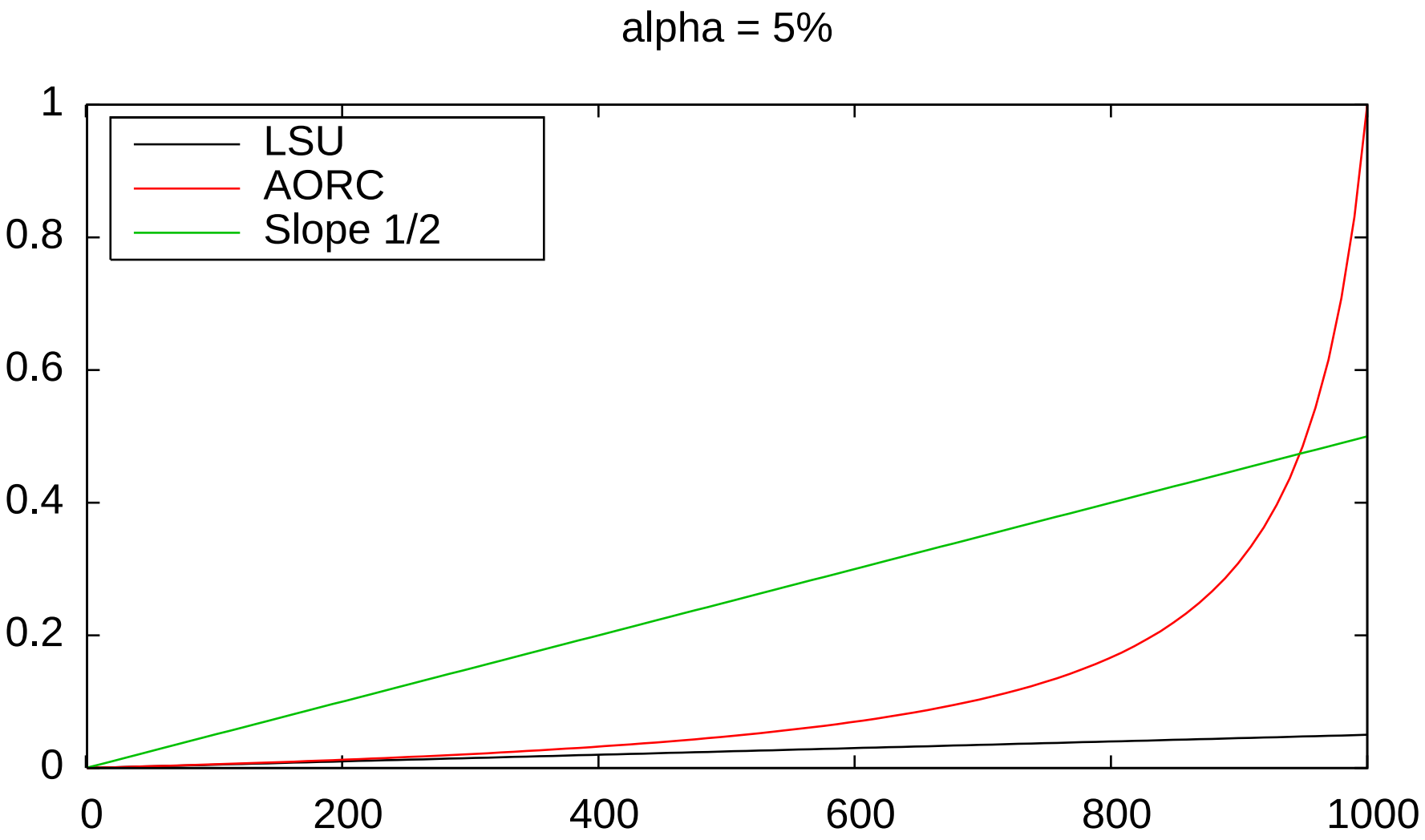
$$\alpha_i = \alpha \frac{i}{m - (1 - \alpha)i} .$$

- problème : comme $\alpha_m = 1$, la procédure step-up utilisant ces seuils rejette systématiquement toutes les hypothèses.

Modifications de la courbe (Finner et al.)

- pour “réparer” la procédure précédente, on peut plafonner l’estimateur sous-jacent de π_0 à une certaine valeur $\pi_0^- > \alpha$.
- autre possibilités : considérer une procédure step up-down.

La courbe de réjection asymptotique de Finner et al.



Analyse asymptotique de Finner et al.

Theorem ([Finner et al.(2009)])

Lorsque le nombre d'hypothèses nulles à tester $|\mathcal{H}| = m$ tend vers ∞ et

- *on a* **(I)** *: les p-values sont indépendantes ;*
- *la proportion $\pi_0(m) = m_0/m$ tend vers une constante π_0 lorsque $m \to \infty$.*

Alors les différentes modifications de la courbe de réjection optimale sont telles que

$$\limsup_m FDR(R_m) \leq \alpha \, .$$

Éléments de preuve

- par Glivenko-Cantelli, pour tout $\lambda_0 < 1$,

$$\limsup_{m\to\infty} \sup_{\lambda\leq\lambda_0} \widehat{\pi}_0^{Storey}(\mathbf{p}_m, \lambda) = \limsup_{m\to\infty} \sup_{\lambda\leq\lambda_0} \frac{|\{h : p_h \leq \lambda\}|}{(1-\lambda)m}$$
$$= \sup_{\lambda\leq\lambda_0} \frac{F_\infty(\lambda)}{(1-\lambda)} \leq \pi_0 .$$

- il est donc asymptotiquement valide d'utiliser l'estimateur de Storey-λ avec un λ aléatoire pourvu que $\lambda \leq \lambda_0 < 1$.
- les modifications de la courbe de réjection optimale assurent que cette condition est remplie (à nouveau par Glivenko-Cantelli)

Une approximation non-asymptotique de l'AORC

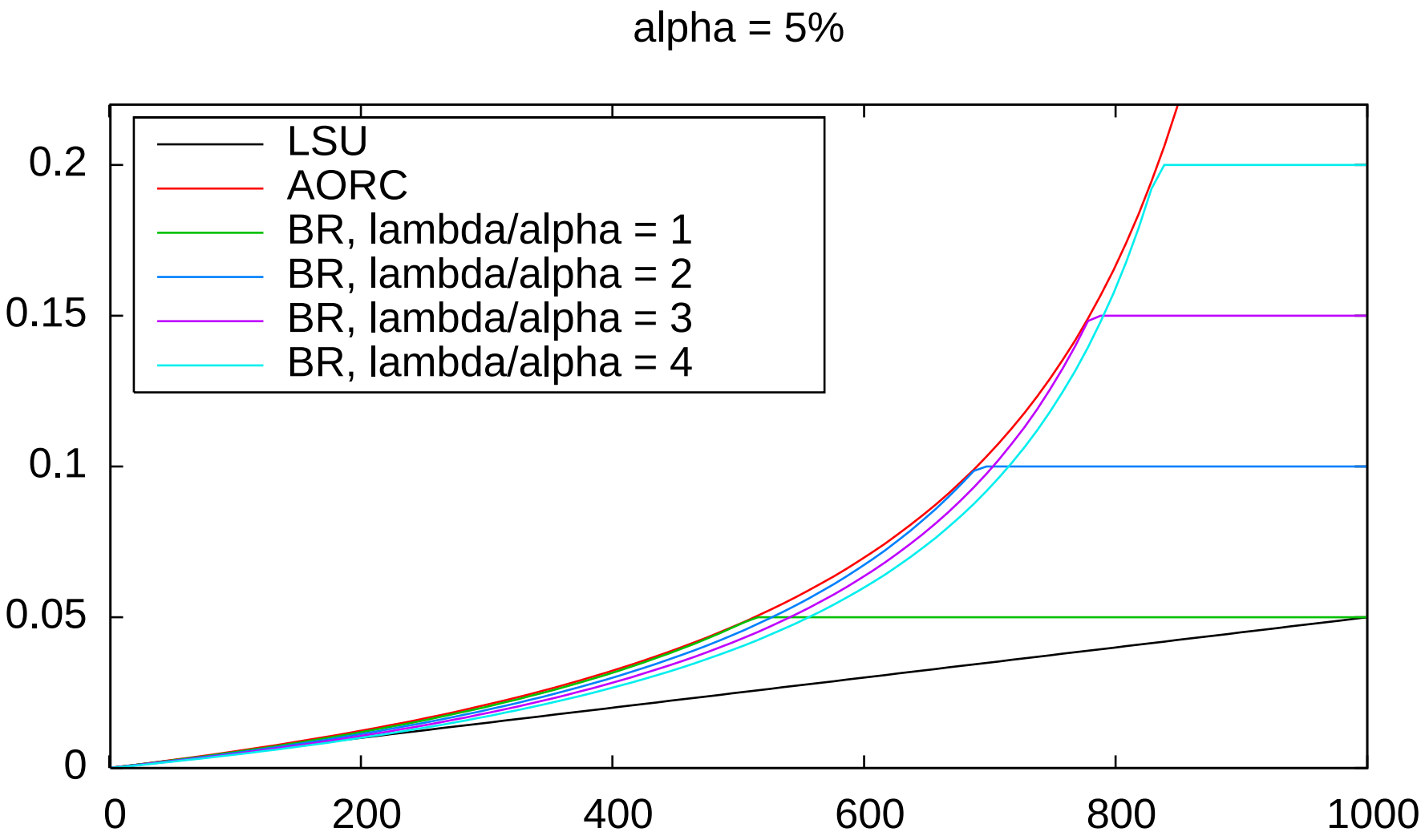
Theorem

On suppose l'hypothèse **(I)** *vérifiée. Soit* $\lambda \in (0,1)$ *. Alors, toute procédure self-consistante par rapport aux seuils*

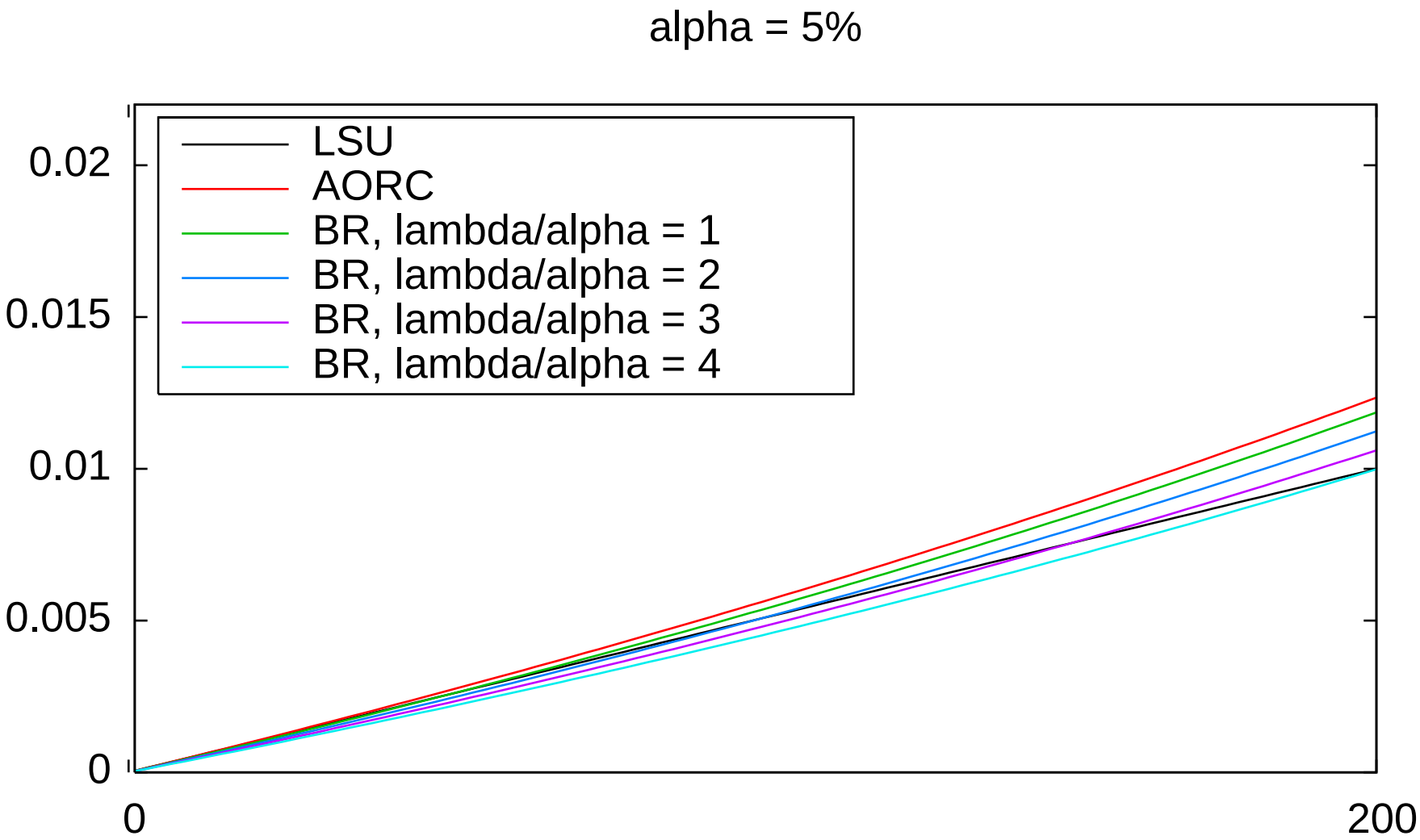
$$\alpha_i = \min\left((1-\lambda)\frac{\alpha i}{m-i+1}, \lambda\right)$$

a un FDR borné par α *pour tout* $|\mathcal{H}| = m$ *fini.*

Comparaison des courbes de réjection



Comparaison des courbes de réjection



FDR et adaptivité à π_0 : résumé

- On peut utliser des nombreux estimateurs de la proportion π_0 , (par ex. Meinshausen et Rice 2006)
- On peut utiliser des inégalités sur les déviations de cet estimateur, ou sur l'espérance de son inverse (sou **(I)**) pour obtenir des procédures adaptatives avec garantie sur le FDR.
- Les procédure adaptatives ont seulement un intérêt si on s'attend à ce que π_0 soit significativement diffrent de 1. Pour des situations "sparse", il n'y pas d'amélioration à attendre sur les procédure de base.
- Un choix "data-dependent" du paramètre λ dans l'estimateur de Storey donne lieu a des procédures adaptives "en une étape" où l'estimation de π_0 est implicite
- Pour une analyse fine des propriétés asymptotiques de convergence (type TCL) sous **(I)**, voir aussi Neuvial (2008).

Et en pratique ?

- les différerentes procédures adaptatives proposées ont été étudiées théoriquement sous l'hypothèse**(I)**.
- les procédures adptatives peuvent être plus ou moins fragiles par rapport à des violations de l'hypothèse **(I)**.
- dans des simulations extensives sur des *p*-values positivement corrélées, la procédure en 2 étapes utilisant l'estimateur de Storey avec $\lambda = \alpha$ est apparue la plus robuste (au prix d'une certaine conservativité quand **(I)** est vérifiée)

Et en pratique ?

- les différerentes procédures adaptatives proposées ont été étudiées théoriquement sous l'hypothèse**(I)**.
- les procédures adptatives peuvent être plus ou moins fragiles par rapport à des violations de l'hypothèse **(I)**.
- dans des simulations extensives sur des *p*-values positivement corrélées, la procédure en 2 étapes utilisant l'estimateur de Storey avec $\lambda = \alpha$ est apparue la plus robuste (au prix d'une certaine conservativité quand **(I)** est vérifiée)

Et en pratique ?

- les différerentes procédures adaptatives proposées ont été étudiées théoriquement sous l'hypothèse**(I)**.
- les procédures adptatives peuvent être plus ou moins fragiles par rapport à des violations de l'hypothèse **(I)**.
- dans des simulations extensives sur des *p*-values positivement corrélées, la procédure en 2 étapes utilisant l'estimateur de Storey avec $\lambda = \alpha$ est apparue la plus robuste (au prix d'une certaine conservativité quand **(I)** est vérifiée)

Plan

1. Contrôle du FDR : conditions suffisantes
2. Adaptativité à π_0
3. Relation avec le problème de la classification

Motivation

- Rappelons le modèle fréquemment utilisé **(RE)** :
- $h_1, \ldots, h_N$ des variables de Bernoulli indépendantes de paramètre $1 - \pi_0$, les p-values sont indépendantes conditionnellement aux (h_i) avec

$$p_i \sim \begin{cases} U([0,1]) & \text{si } h_i = 0\,, \\ P_1 & \text{si } h_i = 1\,. \end{cases}$$

- Ainsi les p-values suivent un modèle de mélange $\pi_0 U + (1 - \pi_0) P_1$, semblable à un modèle de classification.

Motivation

- En général, on suppose que la fonction génératrice de P_1 est concave, ce qui assure que les ensembles de rejet $R_t = \{p_h \leq t\}$ sont UPP au sens de Neyman-Pearson de U contre P_1 .
- Ce peut ne pas être le cas en pratique. . .
- Sun et Cai (2007) proposent d'estimer la densité de P_1 ainsi que la proportion π_0 afin de les utiliser comme plug-in dans des tests de rapports de vraisemblance.
- Cela suppose qu'on a déjà fixé une statistique de test.
- De plus l'hypothèse que p_h est uniforme implique que l'on connaît parfaitement la loi de l'hypothèse nulle – pas toujours vrai, par ex. en comparaison à un groupe de contrôle.

Détection d'anomalies par rapport à une référence

- Observations sur un espace $\mathcal{X}$
- P_0 distribution de référence ou nominale. P_1 distribution "d'anomalies".
- Observations :
  - un échantillon de référence $X_1, \ldots, X_m \sim P_0$
  - un échantillon "contaminé" $X_{m+1}, \ldots, X_{m+n} \sim P_X = \pi_0 P_0 + (1-\pi_0) P_1$.
- pour un classifieur $f : \mathcal{X} \to \{0, 1\}$ on définit

$$\mathcal{E}_y(f) = P_y(f \neq y)\,.$$

et

$$\mathcal{E}_X(f) = P_X(f \neq 1) = \pi_0(1 - \mathcal{E}_0(f)) + (1-\pi_0)\mathcal{E}_1(f)\,.$$

- étant donnée un modèle de classifieurs $\mathcal{F}$, on se fixe pour but de trouver un classifieur réalisant le taux optimal de faux négatifs sous contrainte du taux de faux positifs :

$$\mathcal{E}^*_{1,\alpha}(\mathcal{F}) := \inf_{f \in \mathcal{F}} \mathcal{E}_1(f)$$
$$\text{t.q. } \mathcal{E}_0(f) \leq \alpha\,.$$

Réduction du problème

- Définissons l'erreur optimale de test de la distribution contaminée P_X contre P_0 :

$$\mathcal{E}^*_{X,\alpha}(\mathcal{F}) := \inf_{f\in\mathcal{F}} \mathcal{E}_X(f)$$
$$\text{t.q. } \mathcal{E}_0(f) \leq \alpha\,.$$

Proposition

Pour tout $f \in \mathcal{F}$, si $\mathcal{E}_0(f) \leq \alpha + \varepsilon$, alors

$$(\mathcal{E}_1(f) - \mathcal{E}^*_{1,\alpha}(\mathcal{F})) \leq (1-\pi_0)^{-1}(\mathcal{E}_X(f) - \mathcal{E}^*_{X,\alpha}(f) + \pi_0\varepsilon)$$

References


Y. Benjamini, A. M. Krieger, and D. Yekutieli.
Adaptive linear step-up procedures that control the false discovery rate.
Biometrika, 93(3) :491–507, 2006.


G. Blanchard and E. Roquain.
Two simple sufficient conditions for fdr control.
Electronic journal of statistics, 2008.


G. Blanchard and E. Roquain.
Adaptive FDR control under independence and dependence.
Preprint, 2008.


H. Finner, T. Dickhaus, and M. Roters.
On the false discovery rate and an asymptotically optimal rejection curve.
Ann. Statist., 2009.


E. Roquain.
Exceptional motifs in heterogeneous sequences. Contributions to theory and methodology of multiple testing.
PhD Thesis, Universtité Paris-Sud, 2007.