Adaptive procedures for FDR control in multiple testing

G. Blanchard[1] F. Fleuret[2] E. Roquain[3]

[1]Fraunhofer FIRST, Berlin, Germany

[2]CVLAB, EPFL, Lausanne, Switzerland

[3]INRA, Jouy-en-Josas, France

Workshop “algorithms in complex systems”, Eindhoven 24-26 Sept. 2007

Plan

1. Multiple testing
   - Introduction
   - The false discovery rate
2. FDR control from a set-output point of view
   - Self-consistency condition
   - Step-up procedures
3. Adaptive procedures
   - Existing procedures
   - New procedures

Plan

Introduction

Single hypothesis testing: classical topic in statistics.

- observe data sample $\mathbf{X} = (X_1, \ldots, X_n)$
- We want to decide (from observed data) whether a certain assumption $\mathcal{H}_0$ (null hypothesis) on the generating distribution is true or false.
- Examples:
  - Is it true that $\mathbb{E}[X] = 0$?
  - Are the variables (X, Y) independent?
  - Is the distribution of X Gaussian?
  - $\ldots$

Introduction

Single hypothesis testing: classical topic in statistics.

- observe data sample $\mathbf{X} = (X_1, \ldots, X_n)$
- We want to decide (from observed data) whether a certain assumption $\mathcal{H}_0$ (null hypothesis) on the generating distribution is true or false.
- Examples:
  - Is it true that $\mathbb{E}[X] = 0$?
  - Are the variables (X, Y) independent?
  - Is the distribution of X Gaussian?
  - ...

Single testing procedure

- Testing procedure:

$$\text{Data } \mathbf{X} = (X_1, \ldots, X_n) \rightarrow \text{Decision } T(\mathbf{X}) \in \{0, 1\}$$

- $T = 0$ means "null hypothesis accepted" and $T = 1$ "null hypothesis rejected"
- Language convention: if the null hypothesis is rejected, we equivalently call it a "positive detection" or "discovery".
- Type I error (or false positive): $T = 1$ while the null hypothesis $\mathcal{H}_0$ is actually true.
- Type II error (or false negative): $T = 0$ while the null hypothesis $\mathcal{H}_0$ is actually true.

Single testing procedure

- Testing procedure:

$$\text{Data } \mathbf{X} = (X_1, \dots, X_n) \to \text{Decision } T(\mathbf{X}) \in \{0, 1\}$$

- $T = 0$ means "null hypothesis accepted" and $T = 1$ "null hypothesis rejected"
- Language convention: if the null hypothesis is rejected, we equivalently call it a "positive detection" or "discovery".
- Type I error (or false positive): $T = 1$ while the null hypothesis $\mathcal{H}_0$ is actually true.
- Type II error (or false negative): $T = 0$ while the null hypothesis $\mathcal{H}_0$ is actually true.

p-values for single hypothesis testing

- Most testing procedures are based on a test statistic $Z(\mathbf{X}) \in \mathbb{R}$
- $T_\alpha(\mathbf{X}) = \mathbf{1}\{Z(\mathbf{X}) \geq t(\alpha)\}$
- threshold $t(\alpha)$ is such that that, if $\mathcal{H}_0$ is true, $\mathbb{P}\left[Z(\mathbf{X}) \geq t(\alpha)\right] \leq \alpha$
- Ensures control of type I error rate at level α.
- The statistic can then be normalized: put

$$p(\mathbf{X}) = t^{-1}(Z(\mathbf{X}));$$

then if $\mathcal{H}_0$ is true, from the above

$$\mathbb{P}\left[p(\mathbf{X}) \leq \alpha\right] \leq \alpha$$

i.e. $p(\mathbf{X})$ is stochastically lower bounded by a uniform random variable in $[0,1]$.
- $p(\mathbf{X})$ is called the p-value function associated to this testing procedure.

p-values for single hypothesis testing

- Most testing procedures are based on a test statistic $Z(\mathbf{X}) \in \mathbb{R}$
- $T_\alpha(\mathbf{X}) = \mathbf{1}\{Z(\mathbf{X}) \geq t(\alpha)\}$
- threshold $t(\alpha)$ is such that that, if $\mathcal{H}_0$ is true, $\mathbb{P}\left[Z(\mathbf{X}) \geq t(\alpha)\right] \leq \alpha$
- Ensures control of type I error rate at level α.
- The statistic can then be normalized: put

  $$p(\mathbf{X}) = t^{-1}(Z(\mathbf{X}));$$

  then if $\mathcal{H}_0$ is true, from the above

  $$\mathbb{P}\left[p(\mathbf{X}) \leq \alpha\right] \leq \alpha$$

  i.e. $p(\mathbf{X})$ is stochastically lower bounded by a uniform random variable in $[0, 1]$.
- $p(\mathbf{X})$ is called the p-value function associated to this testing procedure.

Multiple testing

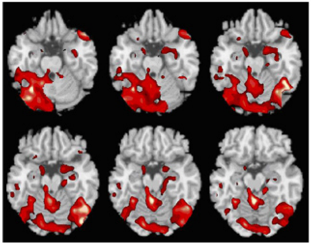
- Possibly very large number of different null hypotheses to test.simultaneously
- testing for the presence of a large number of different chemical compounds.
- testing which pixels represent significant activity in an FMRI image
- testing which genes have significantly high expression level in microarray data
- testing which regression variables $X^{(i)}$ have a dependence relationship with an output Y

Multiple testing

- Possibly very large number of different null hypotheses to test.simultaneously
- testing for the presence of a large number of different chemical compounds.
- testing which pixels represent significant activity in an FMRI image
- testing which genes have significantly high expression level in microarray data
- testing which regression variables $X^{(i)}$ have a dependence relationship with an output Y

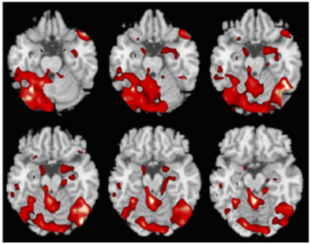
Multiple testing

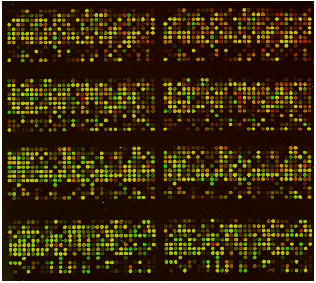
- Possibly very large number of different null hypotheses to test.simultaneously
- testing for the presence of a large number of different chemical compounds.
- testing which pixels represent significant activity in an FMRI image
- testing which genes have significantly high expression level in microarray data
- testing which regression variables $X^{(i)}$ have a dependence relationship with an output Y



Multiple testing

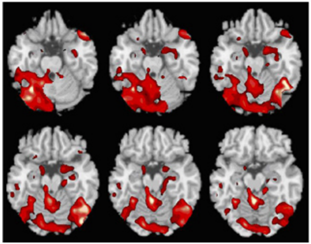
- Possibly very large number of different null hypotheses to test.simultaneously
- testing for the presence of a large number of different chemical compounds.
- testing which pixels represent significant activity in an FMRI image
- testing which genes have significantly high expression level in microarray data
- testing which regression variables $X^{(i)}$ have a dependence relationship with an output Y

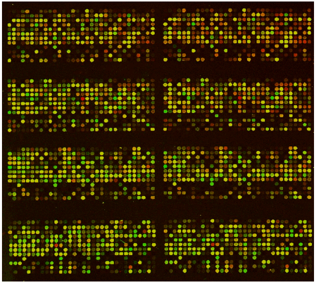




Multiple testing

- Possibly very large number of different null hypotheses to test.simultaneously
- testing for the presence of a large number of different chemical compounds.
- testing which pixels represent significant activity in an FMRI image
- testing which genes have significantly high expression level in microarray data
- testing which regression variables $X^{(i)}$ have a dependence relationship with an output Y





Mathematical setting for multiple testing

- A set $\mathcal{H}$ of null hypotheses to be tested.
- A subset $\mathcal{H}_0 \subset \mathcal{H}$ is the set of null hypotheses that are actually true for the generating probability distribution under scrutiny.
- ($\mathcal{H}_0$ is of course unknown!)
- In general, a multiple testing procedure is:

$$\text{Data } \mathbf{X} = (X_1, \ldots, X_n) \rightarrow \text{Rejected hypotheses } R(\mathbf{X}) \subset \mathcal{H}$$

- We assume that for each single $h \in \mathcal{H}$, we already know a single testing procedure T_h with corresponding p-value function p_h.
- Main issue: how to construct a reasonable multiple testing procedure from the knowledge of the single testing ones?
-

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

Mathematical setting for multiple testing

- A set $\mathcal{H}$ of null hypotheses to be tested.
- A subset $\mathcal{H}_0 \subset \mathcal{H}$ is the set of null hypotheses that are actually true for the generating probability distribution under scrutiny.
- ($\mathcal{H}_0$ is of course unknown!)
- In general, a multiple testing procedure is:

$$\text{Data } \mathbf{X} = (X_1, \ldots, X_n) \rightarrow \text{Rejected hypotheses } R(\mathbf{X}) \subset \mathcal{H}$$

- We assume that for each single $h \in \mathcal{H}$, we already know a single testing procedure T_h with corresponding p-value function p_h.
- Main issue: how to construct a reasonable multiple testing procedure from the knowledge of the single testing ones?
-

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

What quantification for the type I error?

- There is a risk of error for each separate hypothesis tested. How to assess globally the quality of a multiple testing procedure?
- Traditional measure: family-wise error rate (FWER), the probability that the procedure makes at least one type I error:

$$FWER(R) = \mathbb{P}\left[R(\mathbf{X}) \cap \mathcal{H}_0 \neq \emptyset\right]$$

- Less conservative measure of error: Benjamini and Hochberg's False Discovery Rate (FDR) (1995):

$$FDR(R) = \mathbb{E}\left[\frac{|R(\mathbf{X}) \cap \mathcal{H}_0|}{|R(\mathbf{X})|}\right]$$

What quantification for the type I error?

- There is a risk of error for each separate hypothesis tested. How to assess globally the quality of a multiple testing procedure?
- Traditional measure: family-wise error rate (FWER), the probability that the procedure makes at least one type I error:

$$FWER(R) = \mathbb{P}\left[R(\mathbf{X}) \cap \mathcal{H}_0 \neq \emptyset\right]$$

- Less conservative measure of error: Benjamini and Hochberg's False Discovery Rate (FDR) (1995):

$$FDR(R) = \mathbb{E}\left[\frac{|R(\mathbf{X}) \cap \mathcal{H}_0|}{|R(\mathbf{X})|}\right]$$

FDR and screening processes

- The FDR:

$$FDR(R) = \mathbb{E}\left[\frac{|R(\mathbf{X}) \cap \mathcal{H}_0|}{|R(\mathbf{X})|}\right]$$

- This notion is particularly adapted to screening processes:

Large set of objects

$\downarrow$

Select candidates (screening)
using limited data

$\downarrow$

Further study of candidates

Plan

A “self-consistency” condition

- Assume we know *a priori* that $|R| \geq k$.
- If we want $FDR \leq \alpha$, we can afford up to αk errors on average.
- Consider a thresholding procedure $R = \{h : p_h \leq t\}$, then

$$\mathbb{E}\left[\text{Nb of errors for } R\right] = \sum_{h \in \mathcal{H}_0} \mathbb{P}\left[p_h \leq t\right] \leq |\mathcal{H}_0| t \leq |\mathcal{H}| t$$

- Choose $t = \alpha k / |\mathcal{H}|$.
- Now, if for an arbitrary procedure R we observe “post-hoc” that we would like to have rejected $\{h : p_h \leq \alpha |R| / |\mathcal{H}|\}$.
- Introduce the self-consistency condition

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

- β is the shape function.

A “self-consistency” condition

- Assume we know *a priori* that $|R| \geq k$.
- If we want $FDR \leq \alpha$, we can afford up to αk errors on average.
- Consider a thresholding procedure $R = \{h : p_h \leq t\}$, then

$$\mathbb{E}\left[\text{Nb of errors for } R\right] = \sum_{h \in \mathcal{H}_0} \mathbb{P}\left[p_h \leq t\right] \leq |\mathcal{H}_0| t \leq |\mathcal{H}| t$$

- Choose $t = \alpha k / |\mathcal{H}|$.
- Now, if for an arbitrary procedure R we observe “post-hoc” that we would like to have rejected $\{h : p_h \leq \alpha |R| / |\mathcal{H}|\}$.
- Introduce the self-consistency condition

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

- β is the shape function.

FDR control under **(SC)**

The FDR can rewritten as:

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

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

Under **(SC)** FDR control is reduced to a purely probabilistic bound of the form

$$\mathbb{E}\left[\frac{\mathbf{1}\{U \leq c\beta(V)\}}{V}\right] \leq c$$

where U is stochastically lower bounded by a uniform distribution, and under appropriate dependency conditions between U and V.

Different cases

- Case 1: independent test statistics: inequality satisfied for $\beta(x) = x/|\mathcal{H}|$.
- Case 2: positively dependent (PRDS, Benjamini and Yekutieli 2001) test statistics: inequality satisfied for $\beta(x) = x/|\mathcal{H}|$.
- Case 3: unspecified dependences. inequality satisfied for

$$\beta(x) = \frac{1}{|\mathcal{H}|} \int_0^x u d\nu(u) \, ,$$

where ν is any probability measure on $\mathbb{R}_+$.

In all of these three cases: under the corresponding dependency assumptions and **(SC)**, we have

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

In cases 1 and 2: additionally assume that R is a nonincreasing function of **p**.

Different cases

- Case 1: independent test statistics: inequality satisfied for $\beta(x) = x/|\mathcal{H}|$.
- Case 2: positively dependent (PRDS, Benjamini and Yekutieli 2001) test statistics: inequality satisfied for $\beta(x) = x/|\mathcal{H}|$.
- Case 3: unspecified dependences. inequality satisfied for

$$\beta(x) = \frac{1}{|\mathcal{H}|} \int_0^x u d\nu(u) \,,$$

where ν is any probability measure on $\mathbb{R}_+$.

In all of these three cases: under the corresponding dependency assumptions and **(SC)**, we have

$$FDR(R) \le \frac{|\mathcal{H}_0|}{|\mathcal{H}|}\alpha$$

In cases 1 and 2: additionally assume that R is a nonincreasing function of **p**.

Different cases

- Case 1: independent test statistics: inequality satisfied for $\beta(x) = x/|\mathcal{H}|$.
- Case 2: positively dependent (PRDS, Benjamini and Yekutieli 2001) test statistics: inequality satisfied for $\beta(x) = x/|\mathcal{H}|$.
- Case 3: unspecified dependences. inequality satisfied for

$$\beta(x) = \frac{1}{|\mathcal{H}|} \int_0^x u d\nu(u) \,,$$

  where ν is any probability measure on $\mathbb{R}_+$.

In all of these three cases: under the corresponding dependency assumptions and **(SC)**, we have

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

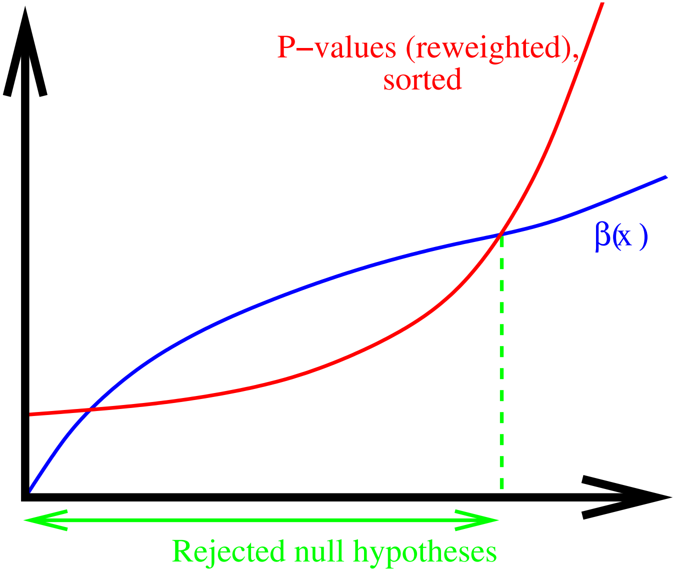
In cases 1 and 2: additionally assume that R is a nonincreasing function of $\mathbf{p}$

Step-up procedures

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

- any procedure R satisfying **(SC)** for a certain shape function β has controlled FDR under appropriate dependency conditions.
- to optimize power under this constraint, we want to have the set R as large as possible under **(SC)**.
- this is precisely realized by a "step-up" procedure:
  - order the p-values $p^{(1)} \leq p^{(2)} \leq \ldots \leq p^{(m)}$
  - put $\widehat{k} = \max\{i : p^{(i)} \leq \alpha\beta(i)\}$
  - put $R = \left\{h^{(1)}, \ldots, h^{(\widehat{k})}\right\}$

Step-up procedure



Role of shape function β

- In the case of independent or PRDS test statistics, β is linear with slope $|\mathcal{H}|^{-1}$. It is the celebrated Benjamini-Hochberg (1995) linear step-up procedure (LSU)
- In the case of unspecified dependencies, we have to pay a price: shape function β is always smaller than the LSU.

$$\beta(x) = |\mathcal{H}|^{-1} \int_0^x u d\nu(u)$$

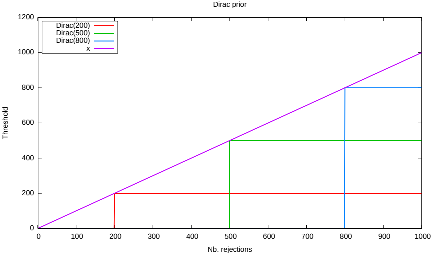
- (Counter-examples exists to show that this price is necessary from a theoretical point of view)
- ν then plays the role of a prior on the rejection set size $|R|$.
- $\nu(i) = c^{-1} i^{-1}$ for $i = 1, \ldots, |\mathcal{H}|$ gives rise to another linear step-up procedure namely recovers Benjamini and Yekutieli (2001). The slope is lower by a factor $c \simeq \ln |\mathcal{H}|$.
- Other choices are possible for the ν-prior and allow added flexibility.

Role of shape function β

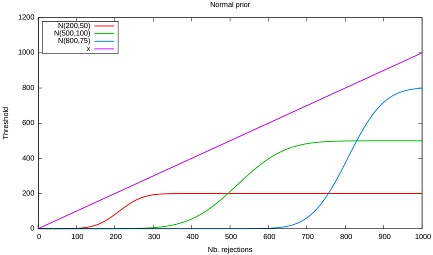
- In the case of independent or PRDS test statistics, β is linear with slope $|\mathcal{H}|^{-1}$. It is the celebrated Benjamini-Hochberg (1995) linear step-up procedure (LSU)
- In the case of unspecified dependencies, we have to pay a price: shape function β is always smaller than the LSU.

$$\beta(x) = |\mathcal{H}|^{-1} \int_0^x u d\nu(u)$$

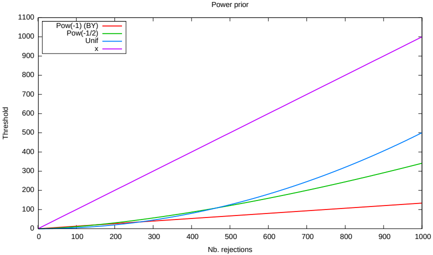
- (Counter-examples exists to show that this price is necessary from a theoretical point of view)
- ν then plays the role of a prior on the rejection set size $|R|$.
- $\nu(i) = c^{-1} i^{-1}$ for $i = 1, \ldots, |\mathcal{H}|$ gives rise to another linear step-up procedure namely recovers Benjamini and Yekutieli (2001). The slope is lower by a factor $c \simeq \ln |\mathcal{H}|$.
- Other choices are possible for the ν-prior and allow added flexibility.



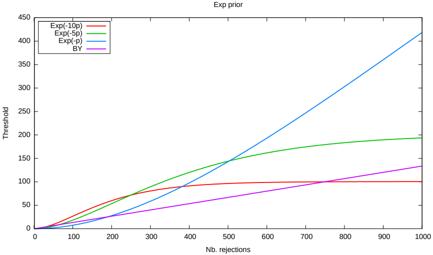
Dirac prior



Normal prior



Power prior



Exponential prior

Plan

Adaptivity to $|\mathcal{H}_0|$

- In all cases reviewed previously, we have derived step-up procedures R satisfying

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

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

- This is always too conservative. Ideally one would replace the shape function β by the "ideal one"

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

- …but π_0 is unknown. Two ways to address this:
  - (Under-)estimate π_0 by some $\widehat{\pi}_0$ then put $\widehat{\beta} = \widehat{\pi}_0^{-1} \beta$ (two-stage procedure).
  - Use a deterministic shape function β that is in some sense directly "adaptive" (one-stage procedure)

Adaptivity to $|\mathcal{H}_0|$

- In all cases reviewed previously, we have derived step-up procedures R satisfying

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

where $\pi_0 = \frac{|\mathcal{H}_0|}{|\mathcal{H}|}$.
- This is always too conservative. Ideally one would replace the shape function β by the "ideal one"

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

- ... but π_0 is unknown. Two ways to address this:
  - (Under-)estimate π_0 by some $\widehat{\pi}_0$ then put $\widehat{\beta} = \widehat{\pi}_0^{-1} \beta$ (two-stage procedure).
  - Use a deterministic shape function β that is in some sense directly "adaptive" (one-stage procedure)

Existing procedures

- Modified Storey's procedure (2001): (Storey-λ) 2-stage procedure with

$$\widehat{\pi}_0^{-1} = \frac{(1-\lambda)|\mathcal{H}|}{|\{h : p_h > \lambda\}| + 1}$$

- Procedure of Benjamini, Kruger and Yekutieli (2006) (BKY06): 2-stage procedure with

$$\widehat{\pi}_0^{-1} = \frac{1}{1+\alpha} \frac{|\mathcal{H}|}{|\mathcal{H}| - |R_0|},$$

  where R_0 is the linear step-up procedure at level $\alpha/(1+\alpha)$.
- The following result holds:

Theorem (Benjamini, Kruger, Yekutieli 06)

If we assume that the test statistics are independent, then for either of the above procedures,

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

New one-stage adaptive procedure for independent test statistics

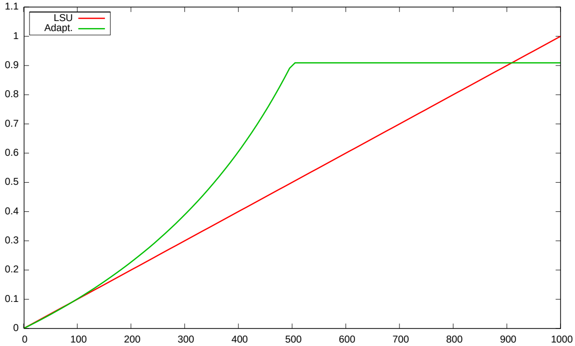
- We introduce a new one-stage step-up procedure:
- Put $\beta(x) = \frac{1}{1+\alpha} \min\left(\frac{x}{|\mathcal{H}|-x+1}, 1\right)$

Theorem

If we assume that the test statistics are independent, then for the one-stage procedure R using the above shape function,

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

Comparison to LSU ($\alpha = 0.1$)



The new one-stage procedure is always more powerful than standard step-up except in "marginal" situations.

New two-stage adaptive procedure for independent test statistics

- Idea: use previous procedure $|R_0'|$ instead of standard linear step-up in (BKY06).
- Use

$$\widehat{\pi}_0^{-1} = \frac{1}{1+\alpha}\frac{|\mathcal{H}|}{|\mathcal{H}| - |R_0'| + 1},$$

Theorem

If we assume that the test statistics are independent, then for the two-stage procedure R using the above $\widehat{\pi}_0^{-1}$,

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

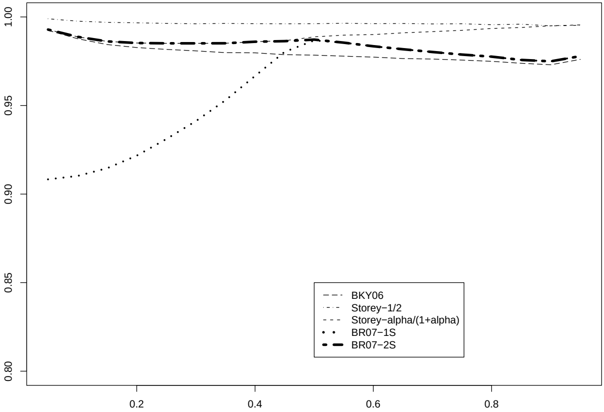
- Always better than (BKY06) except for the "+1" and the marginal situations mentioned previously.

Comparison on simulations

Setting:

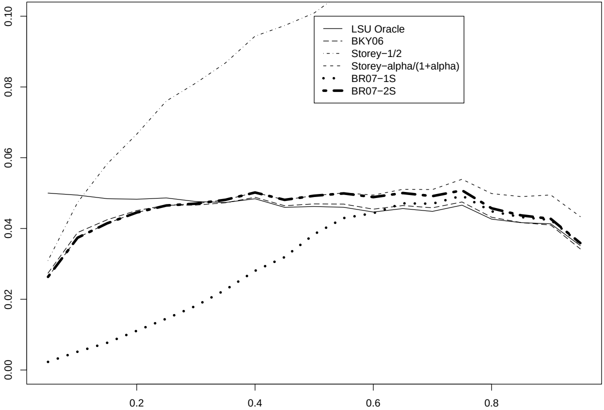
- $X^{(k)} \sim \mathcal{N}(\mu_k, 1)$ with $\mu_k \in \{0, m\}$
- $\mathrm{Cov}\left[X^{(k)}, X^{(k')}\right] = \rho$ for $k \neq k'$
  - $\rho = 0$: independent case
  - $\rho > 0$: positive dependence case
- One-sided tests: $h_k =$ "$\mu_k \leq 0$"
- 1000 repetitions, m=3, $|\mathcal{H}| = 100$

Power (independent case $\rho = 0$)



Plotted: ratio of correct rejections/correct rejections of “oracle” LSU procedure (if π_0 were known), as a function of π_0.

FDR (positive correlation case $\rho = 0.5$



Plotted: FDR against π_0 for various procedures in a positively correlated case

Adaptive procedure under unspecified dependencies

- Recall that for unspecified dependencies we can use a shape function of the form $\beta(x) = \int_0^x u d\nu(u)$.
  - Step 1: perform regular step-up procedure R_0 at level $\alpha/4$ and shape function β.
  - Step 2: put $\widehat{\pi}_0^{-1} = \left(1 - \sqrt{2|R_0|/|\mathcal{H}| - 1}\right)^{-1}$ and perform step-up procedure at level $\alpha/2$ with shape function $\widehat{\beta} = \widehat{\pi_0}^{-1}\beta$

Theorem

The above procedure has $FDR(R) \leq \alpha$ under arbitrary dependencies.

- Note that this is much less favorable than in the independent case. It improves over the non-adaptive procedure only if the first stage at level $\alpha/4$ (!) rejects more than 63% hypotheses.
- However this is up to our knowledge the first theoretically proved adaptive procedure in the unspeficied dependencies case.

Conclusion and perspectives

- Contributions:
  - Synthetic theoretical framework to recover and extend results on FDR control under various dependency assumptions.
  - New one-stage adaptive procededures that improves over standard linear step-up.
  - New two-stage adaptive procedure that improves over (BKY06) and appears robust wrt. positive correlations.
  - In the unspecified dependencies case, first theoretically proved adaptive procedure (relevant only if there is already a large number of "easy" rejections)
- Some orientations for future work:
  - Better adaptive procedures in the unspecified dependencies case
  - Role of the "size prior" ν in the shape function
  - Theoretical support for robustness properties of procedures which are only provably controlled in the independent case