kerfdr: a semi-parametric kernel-based algorithm to Local FDR estimation

M Guedj[1,4], A Célisse[2,4], S Robin[2,4] and G Nuel[3,4]

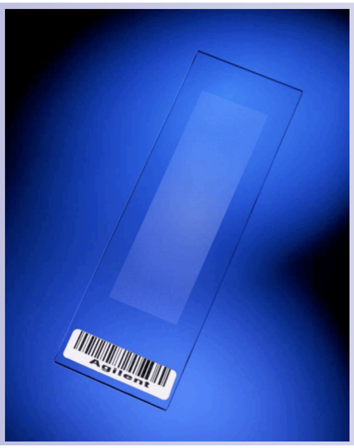
SMPGD 2008, Rennes

[1] Ligue Nationale contre le Cancer, the '*Carte d'Identité des Tumeurs*' group, Paris
[2] Statistics and Genome group, AgroParisTech, INRA, Paris
[3] University Paris Descartes V, MAP5, UMR CNRS 8145, Paris
[4] Statistics for System Biology working group, Paris

Introduction

Thanks to advances in Molecular Biology and improvments of microarray technologies :

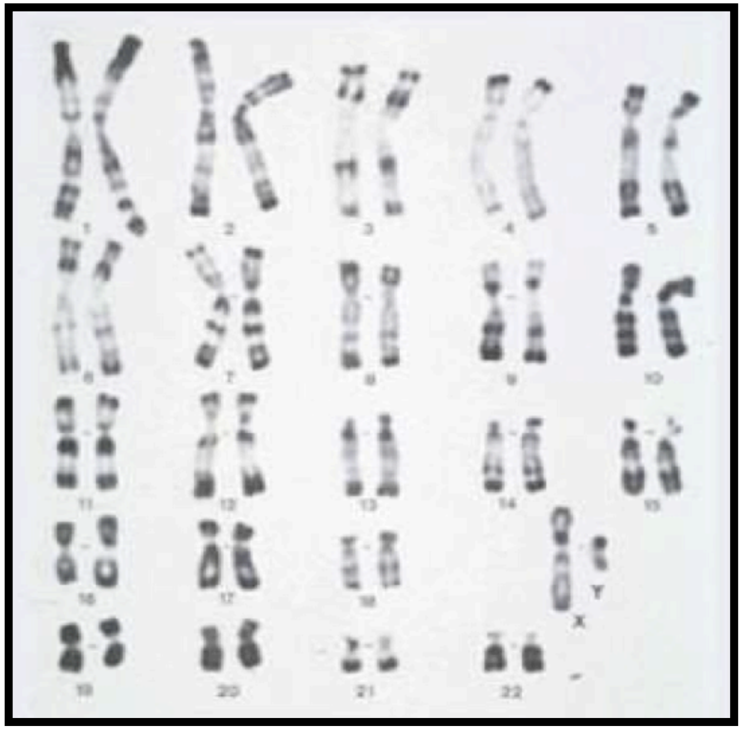
- □ Genome-Wide Associations
- □ Genomic alterations (CGH, CVN)
- □ Gene-Expressions



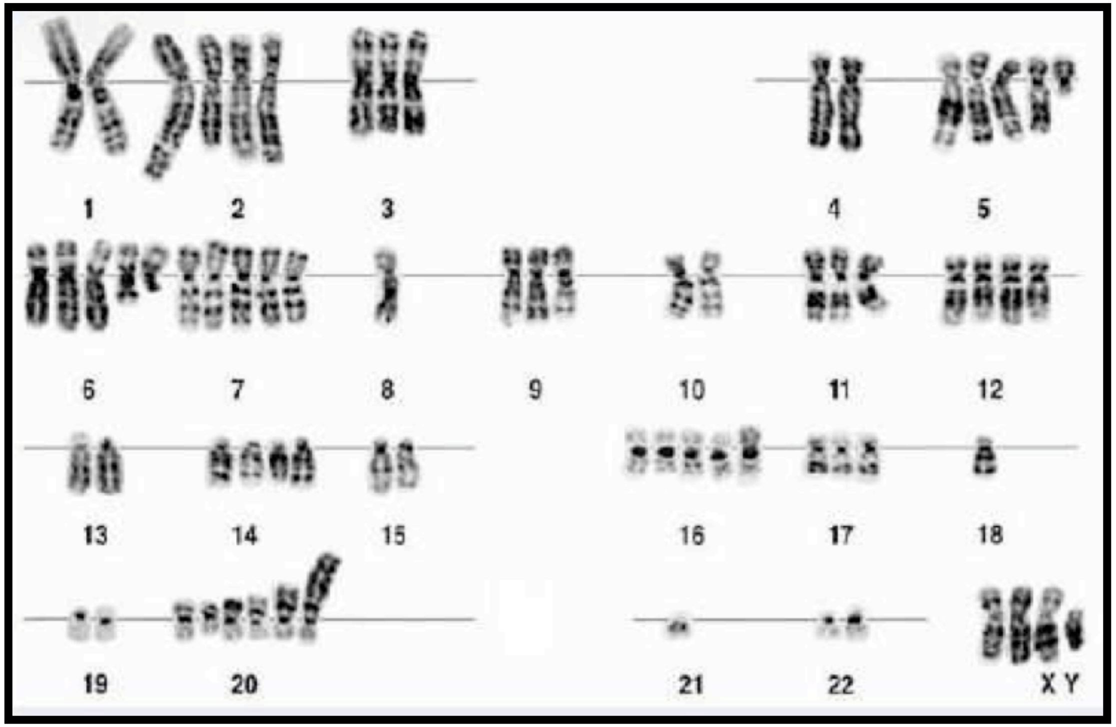
Introduction

- □ Genomic alterations (CGH, CNV):

Normal caryotype



Tumoral caryotype

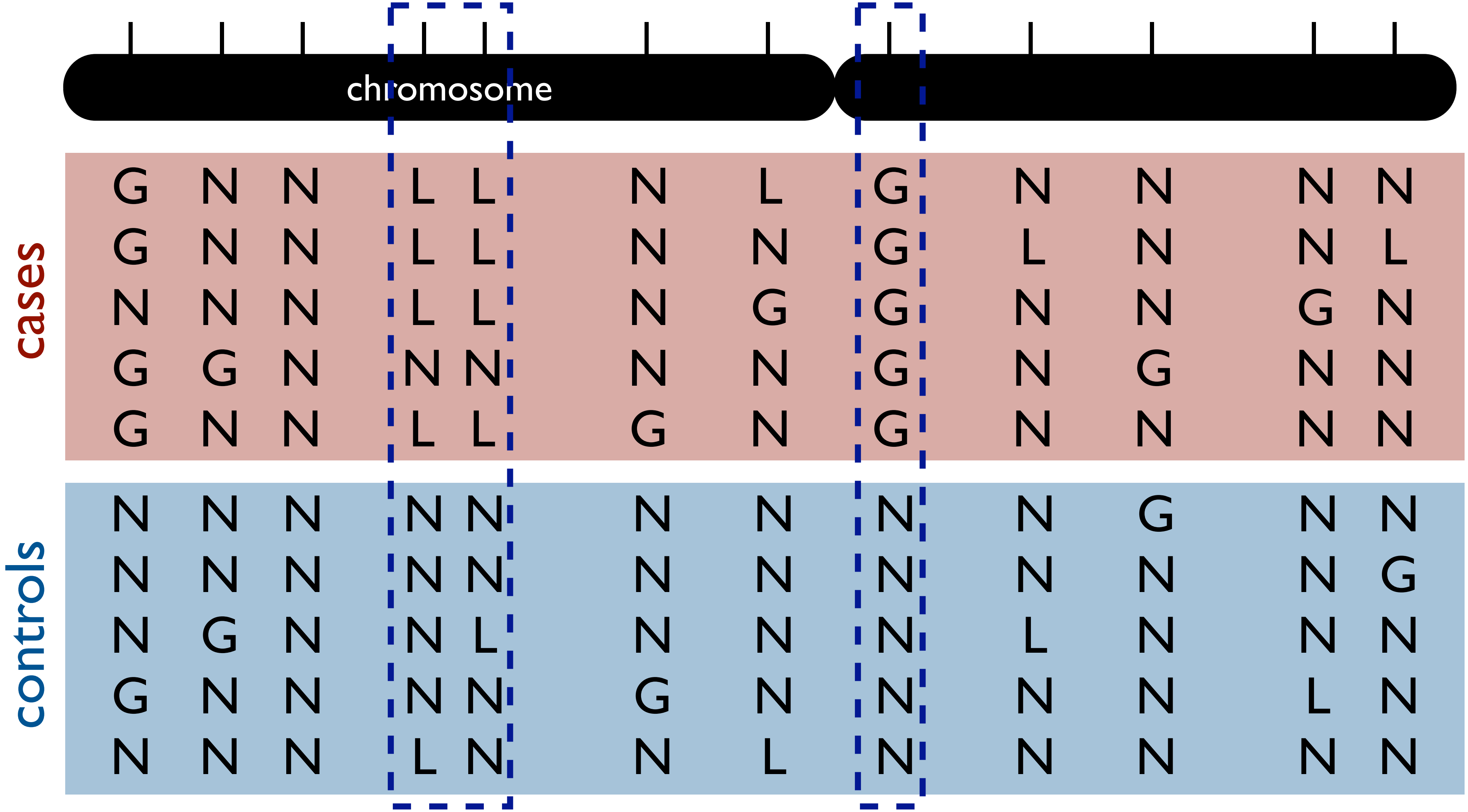


Introduction

L : lost
N : normal
G : gained

□ Genomic alterations (CGH, CNV):

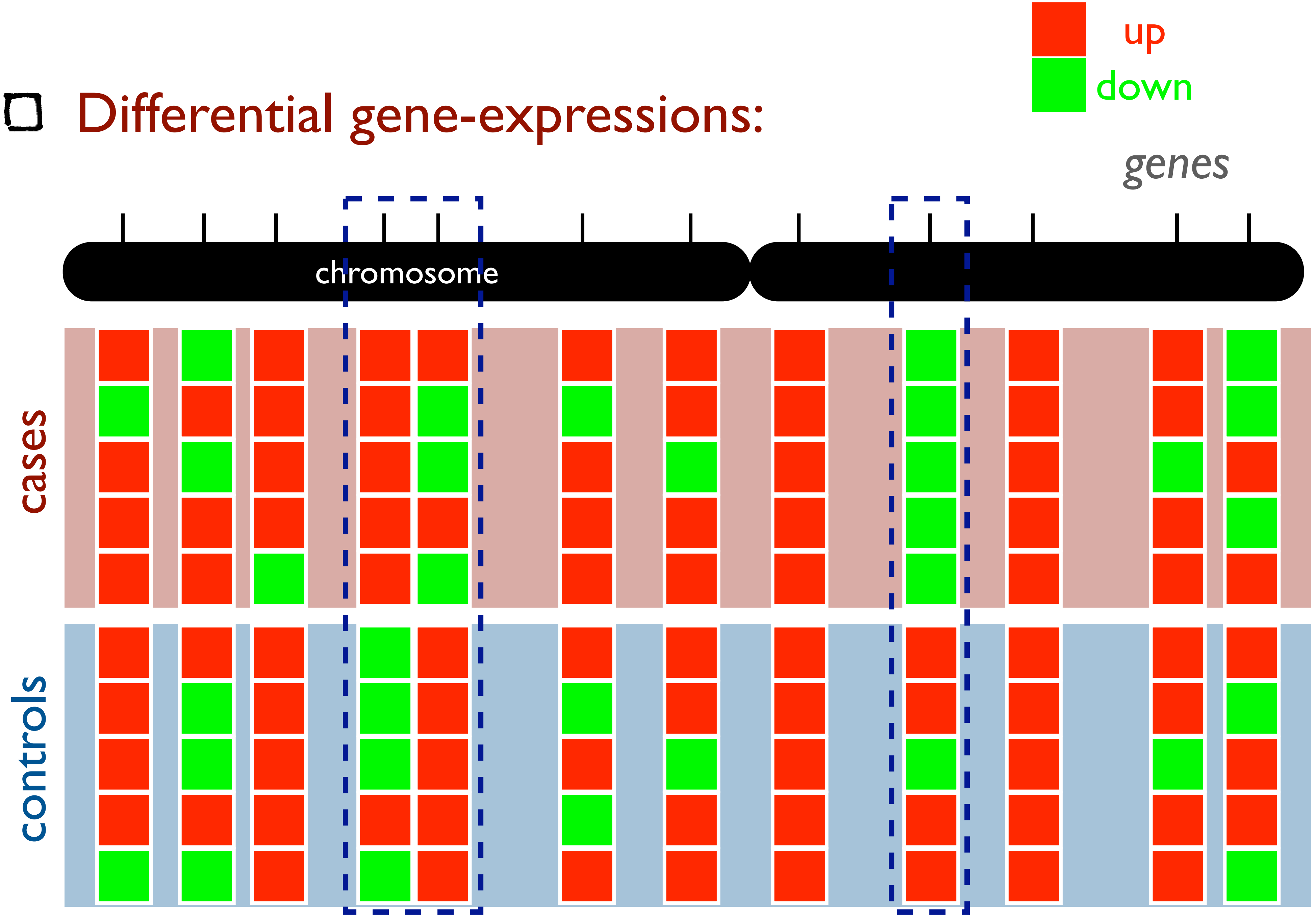
genetic markers



cases	G	N	N	L	L	N	L	G	N	N	N	N
	G	N	N	L	L	N	N	G	L	N	N	L
	N	N	N	L	L	N	G	G	N	N	G	N
	G	G	N	N	N	N	N	G	N	G	N	N
	G	N	N	L	L	G	N	G	N	N	N	N
controls	N	N	N	N	N	N	N	N	N	G	N	N
	N	N	N	N	N	N	N	N	N	N	N	G
	N	G	N	N	L	N	N	N	L	N	N	N
	G	N	N	N	N	G	N	N	N	N	L	N
	N	N	N	L	N	N	L	N	N	N	N	N

Introduction

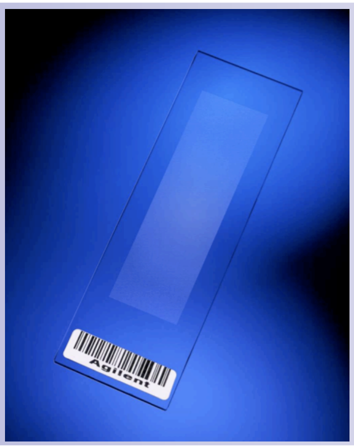
- Differential gene-expressions:



Introduction

Thanks to advances in Molecular Biology and improvments of microarray technologies:

- □ Genome-Wide Associations
- □ Genomic alterations (CGH, CVN)
- □ Gene-Expressions



The use of large-scale data requires the simultaneous evaluation of a huge number of statistical hypotheses.

30,000 genes / 1,000,000 genetic markers (SNPs) ...

- multiple-testing

Introduction

□ n tests at the α level:

	H_0 no rejected	H_0 rejected	
H_0 true	vn	fp	V
H_0 false	fn	vp	F
total	$n - R$	R	n

Introduction

- □ n tests at the α level:

true-negative

	H_0 no rejected	H_0 rejected	
H_0 true	vn	fp	V
H_0 false	fn	vp	F
total	$n-R$	R	n

true-positive

Introduction

- n tests at the α level:

false-positive

	H_0 no rejected	H_0 rejected	
H_0 true	vn	fp	V
H_0 false	fn	vp	F
total	$n - R$	R	n

false-negative

Introduction

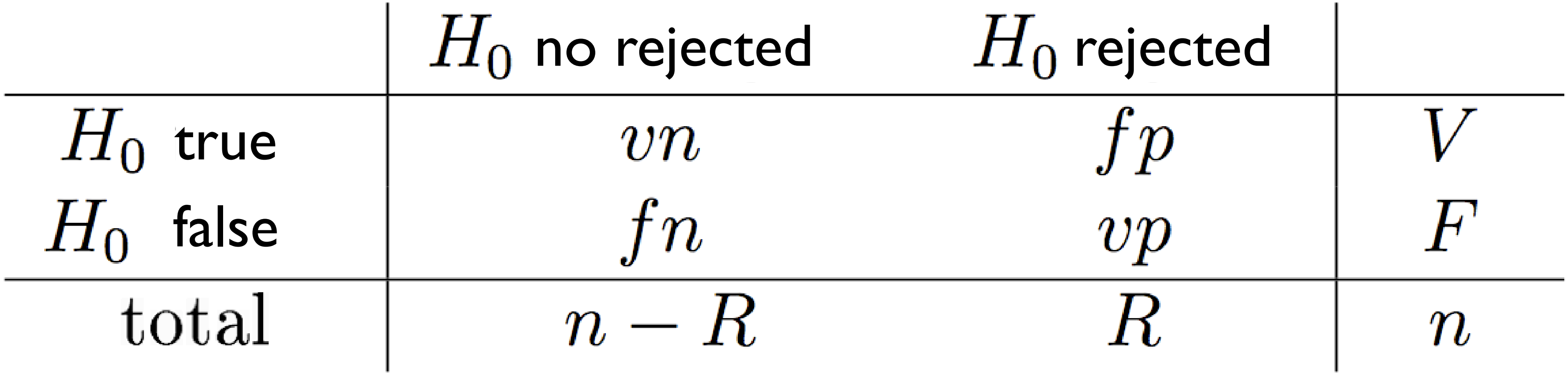
- □ n tests at the α level:

	H_0 no rejected	H_0 rejected	
H_0 true	vn	fp	V
H_0 false	fn	vp	F
total	$n - R$	R	n

- □ $n = 100,000$ $\alpha = 5\%$
  - ▸ 5,000 false-positives >> # true-positives

Introduction

- n tests at the α level:



	H_0 no rejected	H_0 rejected	
H_0 true	vn	fp	V
H_0 false	fn	vp	F
total	$n-R$	R	n

- $n = 100,000 \qquad \alpha = 5\%$
  - 5,000 false-positives >> # true-positives
  - the control of the *fp* is a crucial issue.
  - type-I error-rate not adapted anymore

(Benjamini et Hochberg 95)
(Forner et al 07)

FDR

- less conservative than the FWER
- more intuitive interpretation

□ False Discovery Rate:

$$\mathrm{FDR} = \mathbb{E}(Q),$$

with $Q = \frac{fp}{R}$ if $R > 0$ or $Q = 0$ otherwise.

(Benjamini et Hochberg 95)
(Forner et al 07)

FDR

- less conservative than the FWER
- more intuitive interpretation

❏ False Discovery Rate:

$$\mathrm{FDR} = \mathbb{E}(Q),$$

with $Q = \frac{fp}{R}$ if $R > 0$ or $Q = 0$ otherwise.

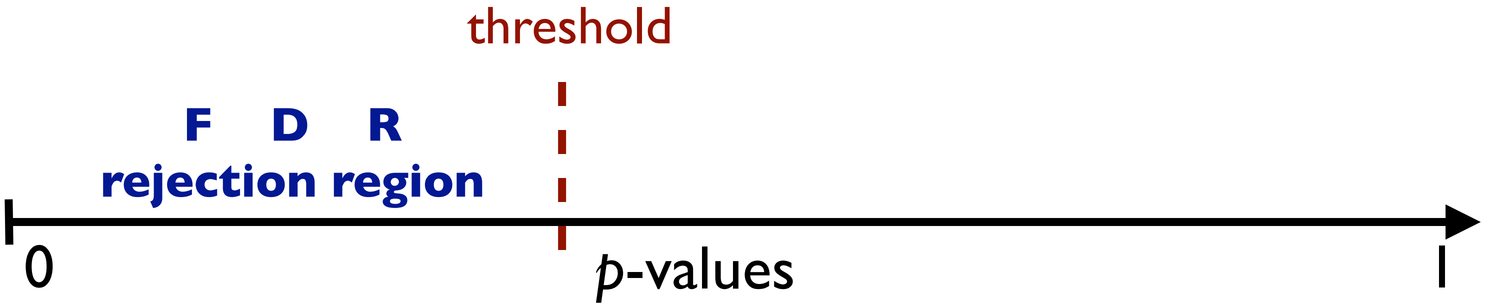
❏ Benjamini-Hochberg's majoration:

$$\mathrm{FDR} \leqslant \ \min \left(\frac{n\alpha}{R(\alpha)}; 1 \right)$$

❏ Estimation with Monte-Carlo simulations.

FDR

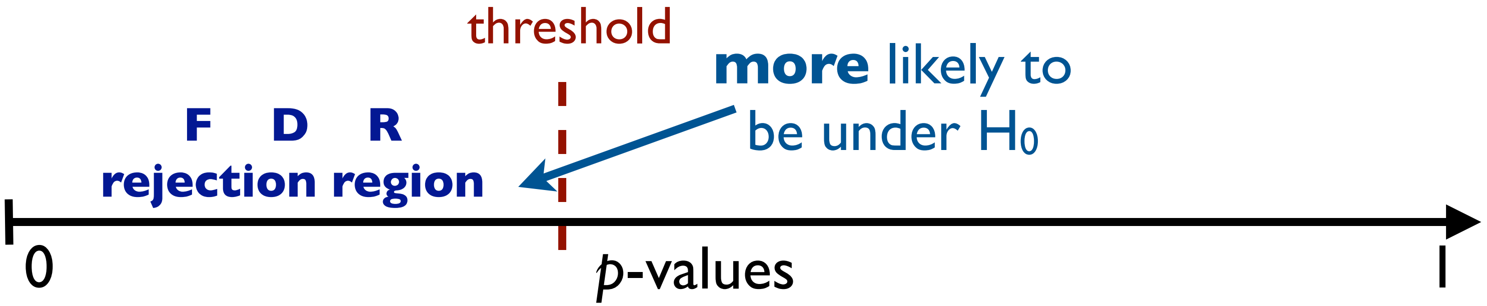
□ False Discovery Rate:



- Global criterion, can not be used to assess the reliability of a specific hypothesis.
- Associated to a given rejection region without distinguishing statistics/p-values that are close to the threshold and those that are not.

FDR

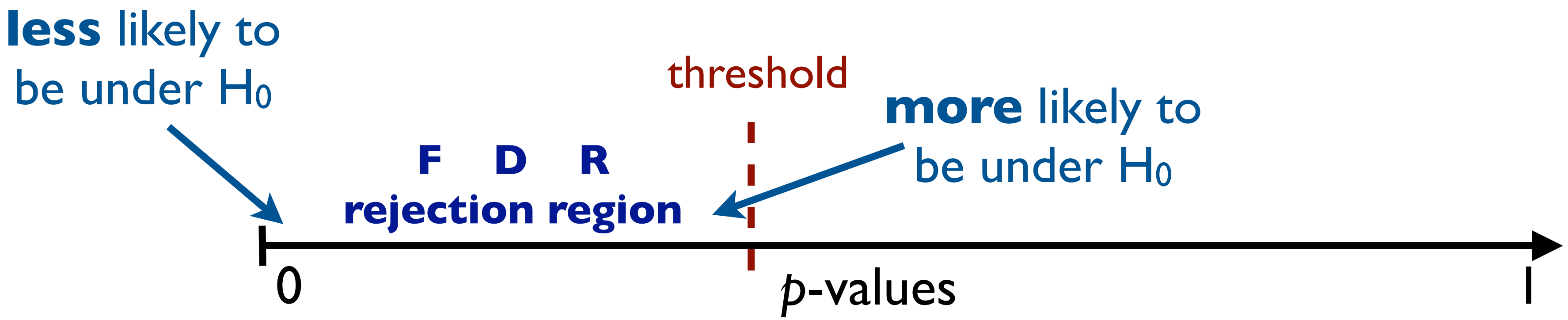
□ False Discovery Rate:



- Global criterion, can not be used to assess the reliability of a specific hypothesis.
- Associated to a given rejection region without distinguishing statistics/*p*-values that are close to the threshold and those that are not.

FDR

□ False Discovery Rate:



less likely to be under H_0

threshold

more likely to be under H_0

F D R

rejection region

0

p-values

1

- **Global criterion,** can not be used to assess the reliability of a specific hypothesis.
- **Associated to a given rejection region** without distinguishing statistics/*p*-values that are close to the threshold and those that are not.

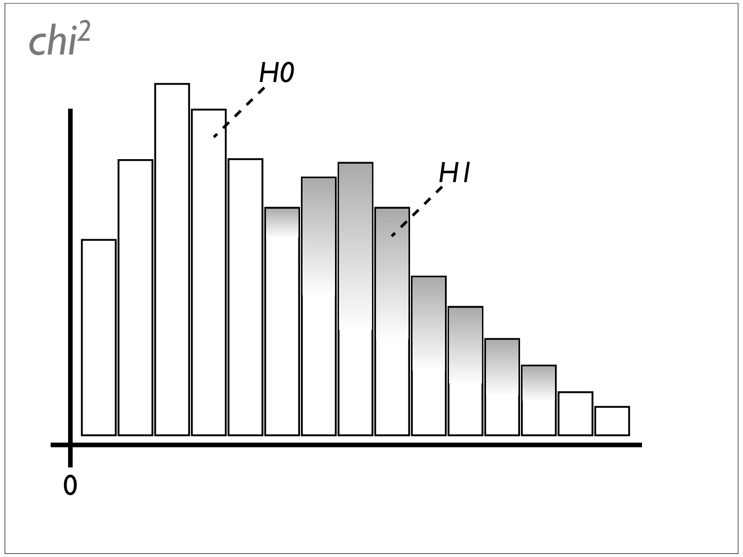
(Efron 04)

Local FDR

□ Local False Discovery Rate:

$$\mathrm{fdr}_i = \mathbb{P}(H = H0 | \mathcal{S} = \mathcal{S}_i)$$

□ Mixture model: general and statistically convenient framework



$$f = \pi_0 f_0 + \pi_1 f_1,$$

$$\mathrm{fdr}_i \equiv \frac{\pi_0 f_0(\mathcal{S}_i)}{f(\mathcal{S}_i)}$$

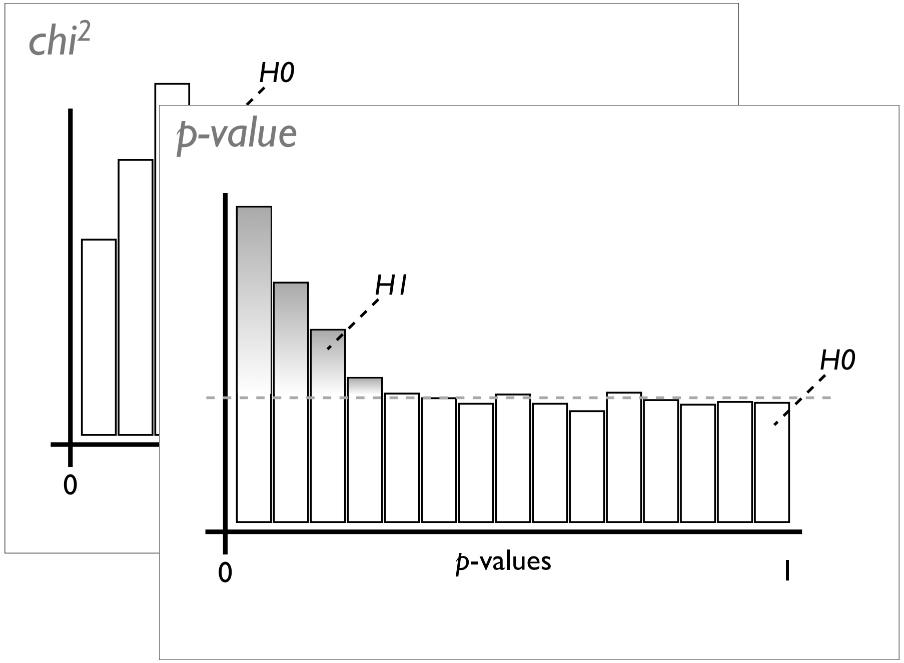
(Efron 04)

Local FDR

- Local False Discovery Rate:

$$\mathrm{fdr}_i = \mathbb{P}\left(H = H0 | \mathcal{S} = \mathcal{S}_i\right)$$

- Mixture model: general and statistically convenient framework



$$f = \pi_0 f_0 + \pi_1 f_1,$$

$$\mathrm{fdr}_i \equiv \frac{\pi_0 f_0(p\mathrm{v}_i)}{f\,(p\mathrm{v}_i)}$$

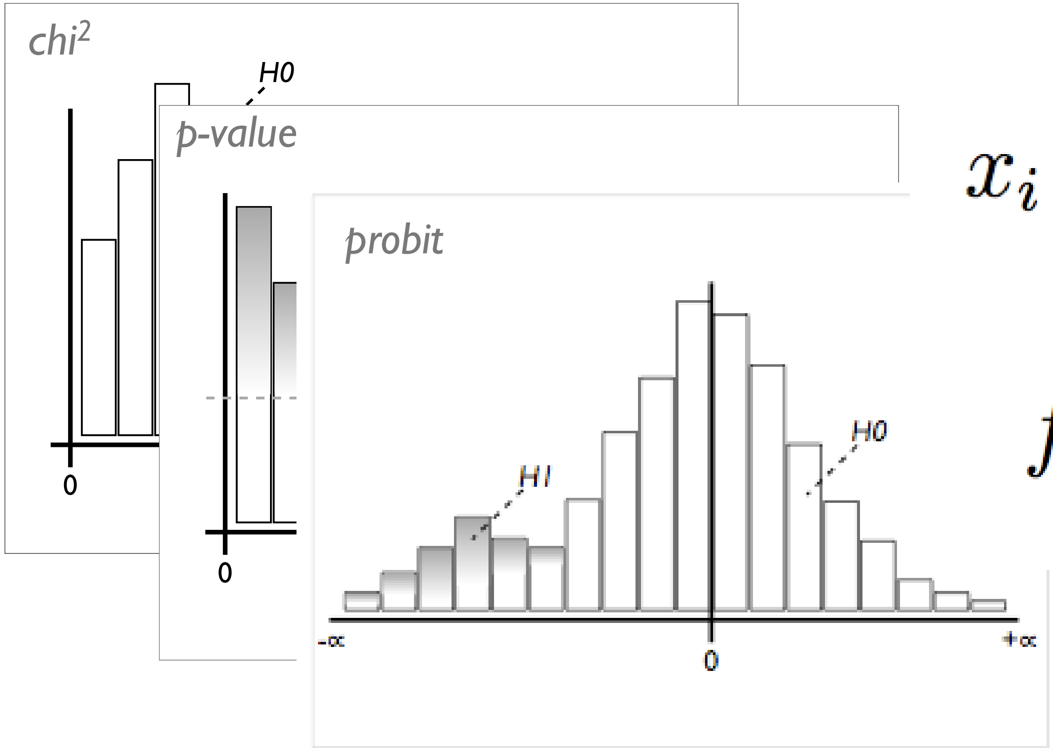
(Efron 04)

Local FDR

- □ Local False Discovery Rate:

$$\mathrm{fdr}_i = \mathbb{P}\left(H = H0 | \mathcal{S} = \mathcal{S}_i\right)$$

- □ Mixture model: general and statistically convenient framework



$$x_i = \mathrm{probit}(pv_i) = \Phi^{-1}(pv_i)$$

$$f_{\theta_j}(x_i) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{\frac{-(x_i - \widehat{\mu}_j)^2}{2(\sigma_j)^2}},$$

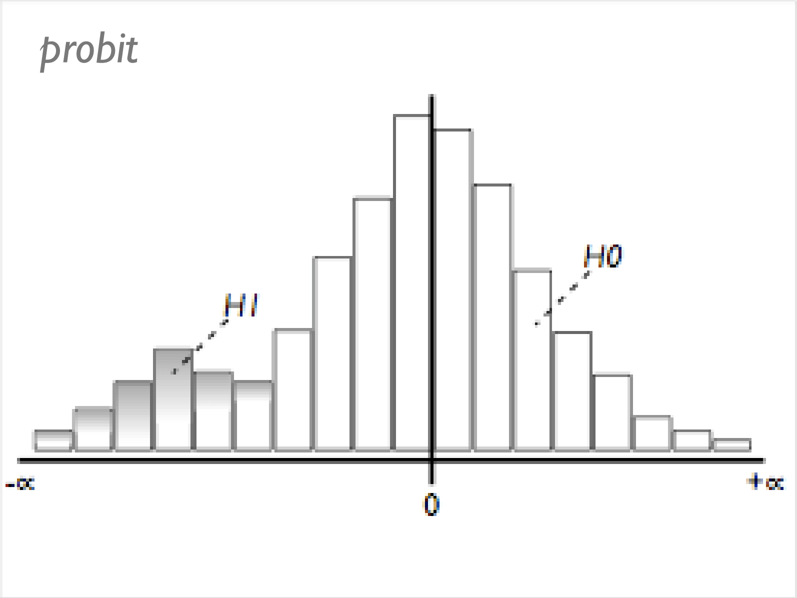
(Efron 04)
(McLachlan et al 06)

Local FDR

- 2-components Gaussian mixture model: EM

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad x_i = \mathrm{probit}(pv_i) = \Phi^{-1}(pv_i),$$

$$\mathrm{fdr}_i \equiv \frac{\pi_0 f_0(x_i)}{f(x_i)} \qquad f_{\theta_j}(x_i) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{\frac{-(x_i - \widehat{\mu}_j)^2}{2(\sigma_j)^2}}$$



$$f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

$$f_1 = \mathcal{N}(\mu_1, \sigma_1)$$

Local FDR

□ 2-components Gaussian mixture model: EM

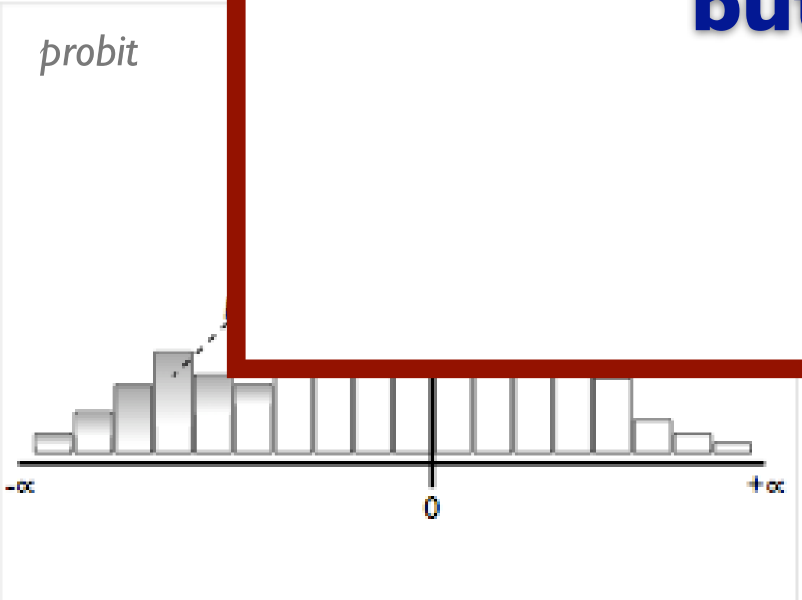
$f =$ $^{1}(pv_i),$

fdr $\frac{(x_i - \hat{\mu}_j)^2}{(\sigma_j)^2}$

Gaussian assumption reasonable

for H_0

but not for H_1



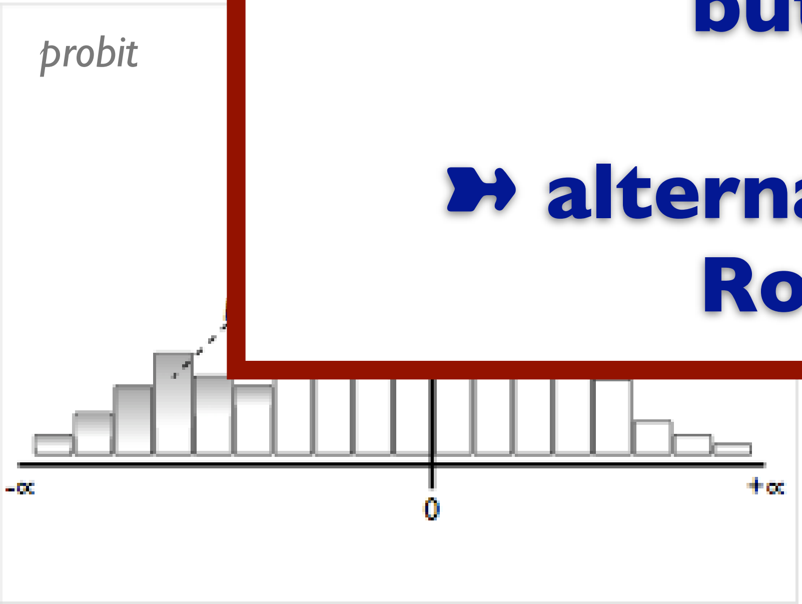
$f_1 \sim \mathcal{N}(\mu_1, \sigma_1)$

Local FDR

- □ 2-components Gaussian mixture model: EM

$f =$ … $^{-1}(pv_i),$

fdr … $\frac{(x_i-\widehat{\mu}_j)^2}{(\sigma_j)^2}$



$f_1 \sim \mathcal{N}(\mu_1, \sigma_1)$

Gaussian assumption reasonable

for H_0

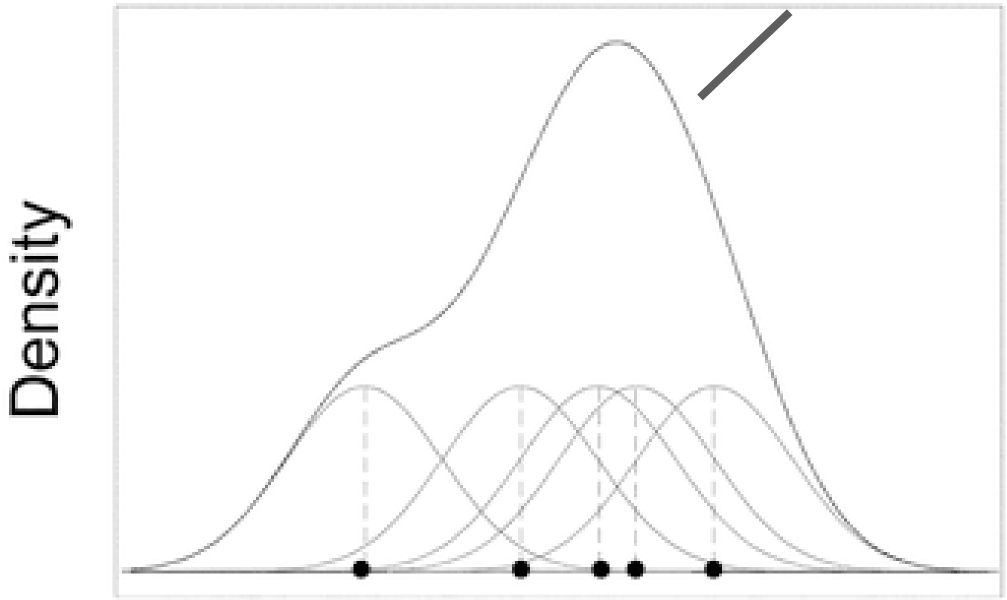
but not for H_1

» alternative proposed by Robin et al 07

kerfdr

□ Kernel-based estimation: non-parametric estimation by convolving the data with a kernel

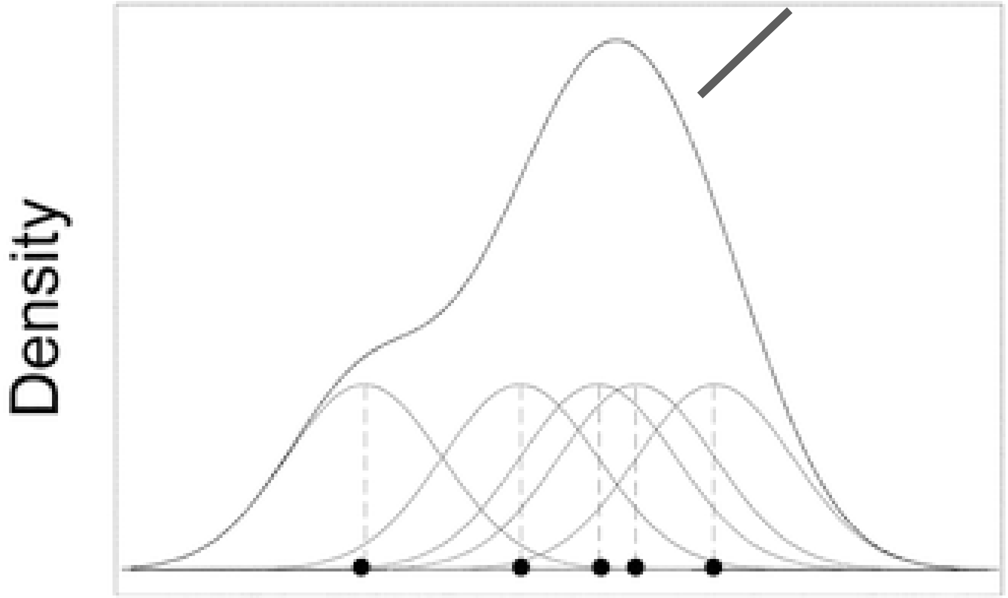
2 parameters



observed density

kerfdr

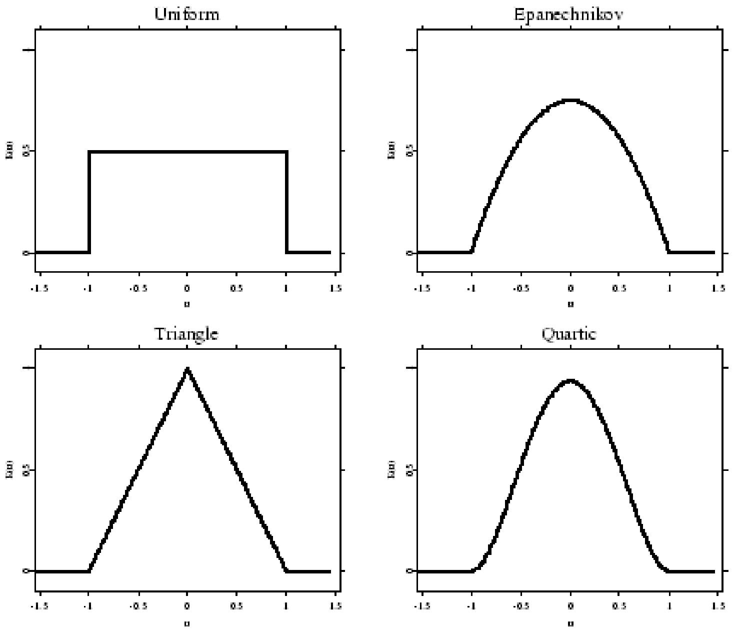
- **Kernel-based estimation:** non-parametric estimation by convolving the data with a kernel



observed density

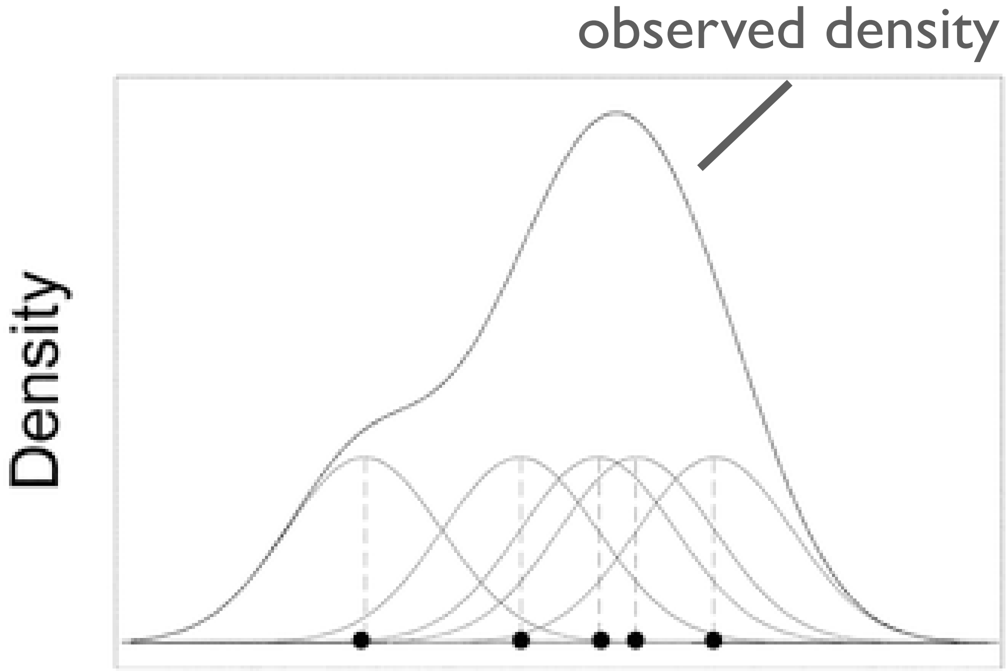
2 parameters

- kernel function (shape)



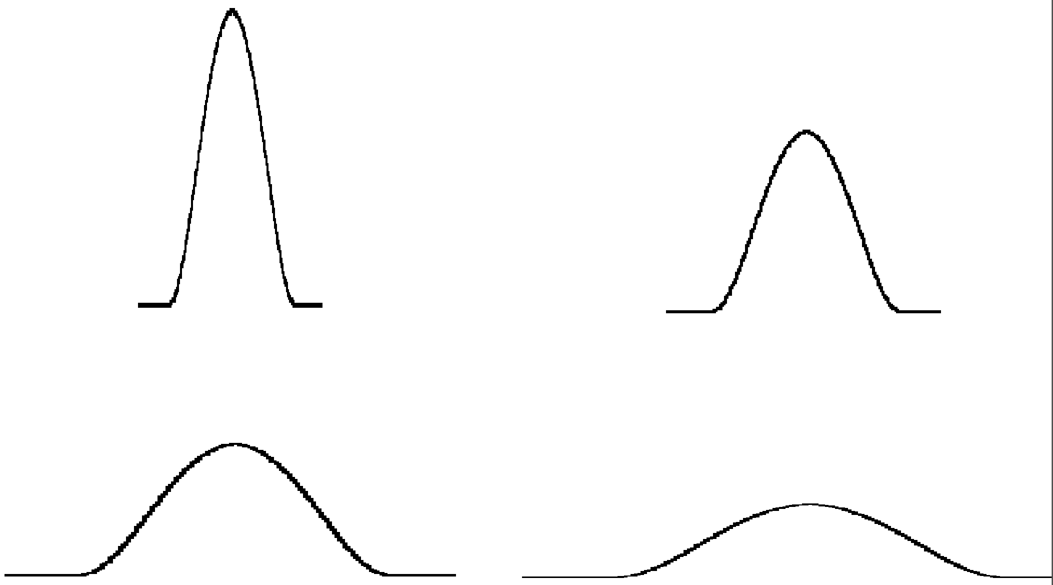
kerfdr

- Kernel-based estimation: non-parametric estimation by convolving the data with a kernel



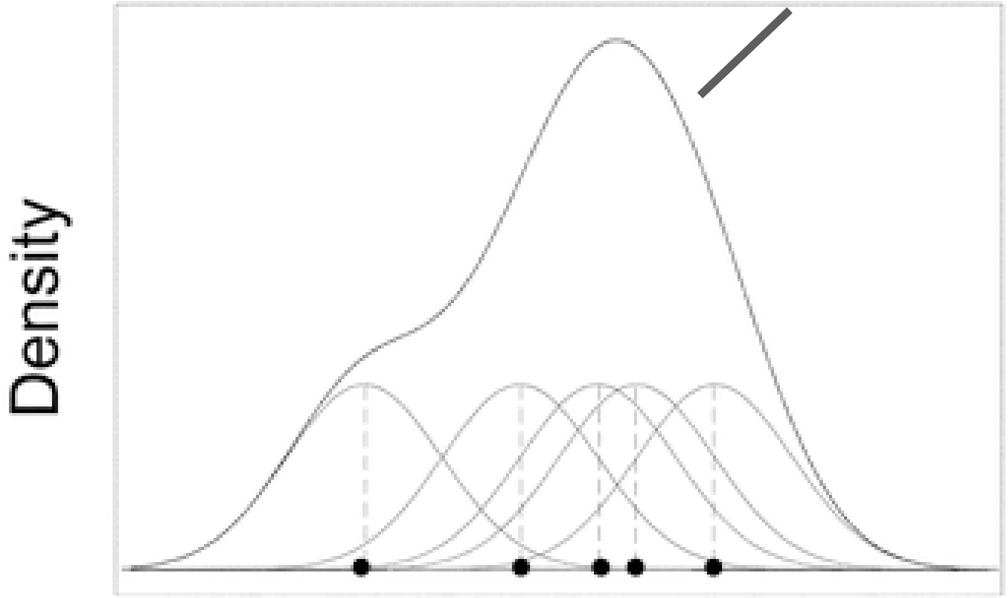
2 parameters

- kernel function (shape)
- bandwidth (smoothing)



kerfdr

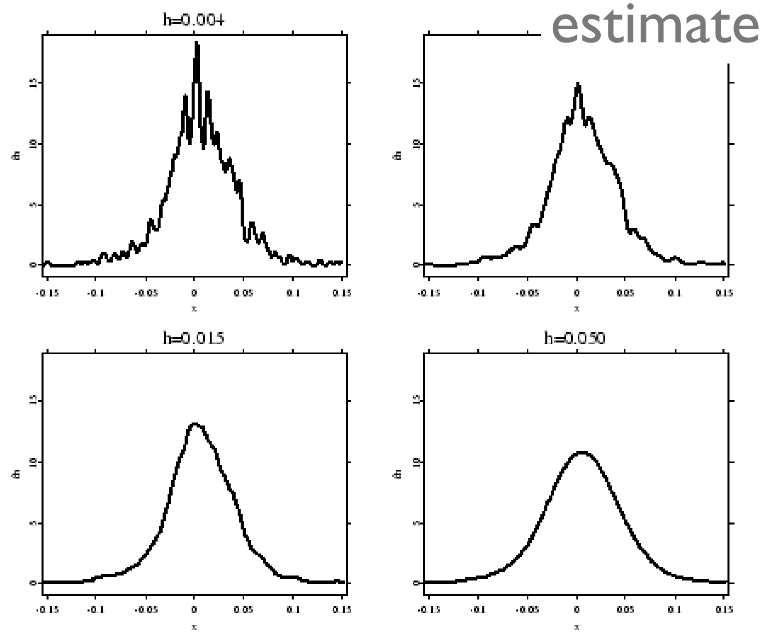
- Kernel-based estimation: non-parametric estimation by convolving the data with a kernel



observed density

2 parameters

- kernel function (shape)
- bandwidth (smoothing)



estimated density

(Robin et al 07)

kerfdr

- Kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

local FDR

kernel function

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i)/\widehat{f}(x_i),$$

bandwidth

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_{i0}}{h} k\left(\frac{x-x_i}{h}\right)\right] \Big/ \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right)$$

(Robin et al 07)

kerfdr

□ Kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i) / \widehat{f}(x_i),$$

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1 - \widehat{\tau}_{i0}}{h} k\left(\frac{x - x_i}{h}\right)\right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right)$$

(Robin et al 07)

kerfdr

- Kernel-based estimation:

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i) / \widehat{f}(x_i),$$

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1 - \widehat{\tau}_{i0}}{h} k \left(\frac{x - x_i}{h} \right) \right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0} \right)$$

(Robin et al 07)

kerfdr

□ Kernel-based estimation: EM-like algorithm

$$f = \pi_0 f_0 + \pi_1 f_1, \qquad f_0 = \mathcal{N}(\mu_0, \sigma_0)$$

Step 'E'

$$\widehat{\tau}_{i0} = \widehat{\pi}_0 f_0(x_i) / \widehat{f}(x_i),$$

Step 'M'

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1 - \widehat{\tau}_{i0}}{h} k \left(\frac{x - x_i}{h} \right) \right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0} \right)$$

kerfdr

- Kernel-based estimation:
  - Semi-parametric.
  - Do not require any assumption on the alternative distribution.
  - Provide more realistic estimates.
  - π_0, h and k must be pre-determined.
  - Tests must be independent.

kerfdr

□ Implementation

- Estimation of π_0
- Determination of the bandwitdh
- Computation of f_l
- Semi-supervised situations
- Truncated distributions

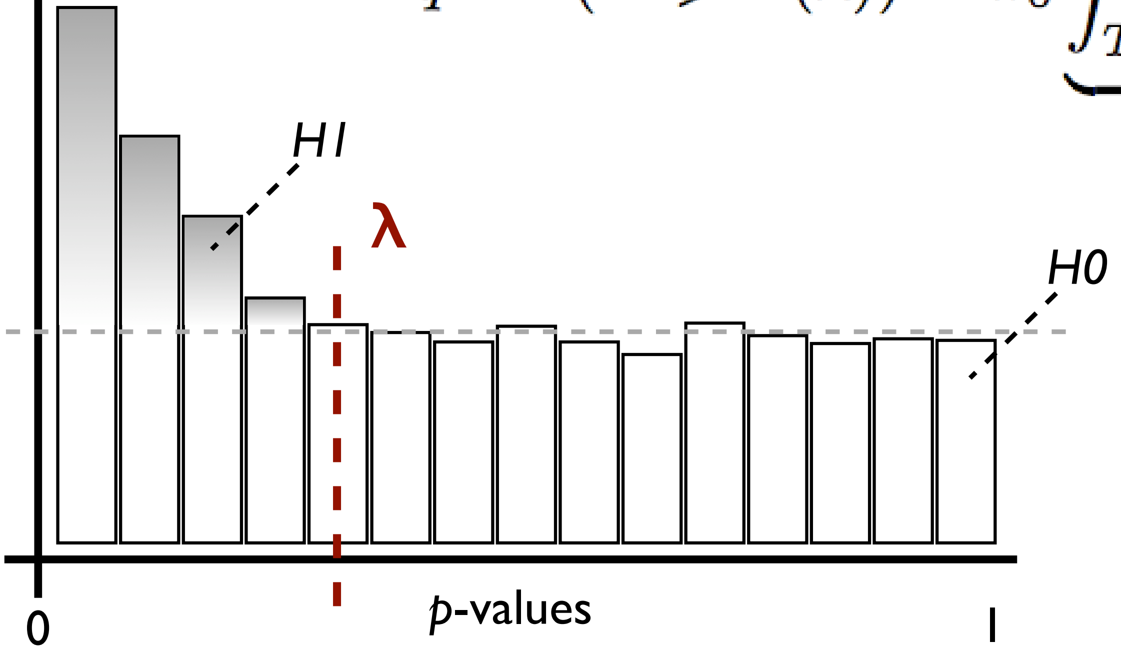
practical generalizations

(Storey 01)

kerfdr

- Implementation
  - Estimation of π_0
- Many methods already implemented

$$q = \mathbb{P}(X \geqslant T(\lambda)) = \pi_0 \underbrace{\int_{T(\lambda)}^{+\infty} f_0(x)dx}_{q_0} + \pi_1 \underbrace{\int_{T(\lambda)}^{+\infty} f_1(x)dx}_{q_1}$$



$$\pi_0 = \frac{q - q_1}{q_0 - q_1}$$

$$\rightarrow \quad \widehat{\pi_0} = \frac{\widehat{q}}{1 - \lambda}$$

(Sheather and Jones 91)
(Silverman 86)
(Scott 92)

kerfdr

- Implementation
- Determination of the bandwidth
- Many methods already implemented :
  - Biased and unbiased cross-validation estimations.
  - Methods using estimation of derivatives.
  - Simple heuristics in the special case of Gaussian kernels.

(Silverman 82)

kerfdr

- Implementation
- Use of Fast Fourier Transforms to compute $\widehat{f}_1(x)$
  - The naive computation requires a quadratic complexity.
  - An algorithm based on fast discrete convolution through FFT allows a far more efficient linear complexity.

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_{i0}}{h} k\left(\frac{x - x_i}{h}\right)\right] / \left(n - \sum_{j=1}^{n} \widehat{\tau}_{j0}\right).$$

kerfdr

- Implementation
- Semi-supervised situations
  - Among the null hypotheses to be tested, some are known to be true (control-genes in dge experiments) while other are known to be false (test genes in spike-in settings).
  - Prior information is taken into account in the estimation procedure.
  - Known local FDR τ_{i0} are kept fixed : they contribute to the estimation for the other observations but are not updated at each step of the algorithm.

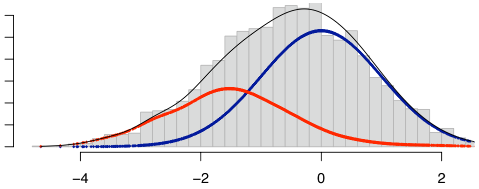
kerfdr

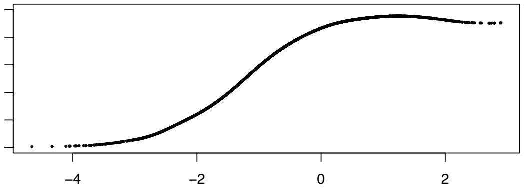
- Implementation
- Truncated distributions within an interval I
  - e.g. : p-values computed by Monte-Carlo ➡ p-values > 1/S
  - the restrictions of f_1, f_0 and f to I need to be normalized with q_1, q_0 and q the corresponding normalization factors.

$$q = \int_I f(x)dx = \pi_0 \underbrace{\int_I f_0(x)dx}_{q_0} + \pi_1 \underbrace{\int_I f_1(x)dx}_{q_1}$$

kerfdr

- Implementation
- R package ‘kerfdr’
  - Simple and straightforward to use
  - Many options for more advanced users
  - Fast thanks to Fast Fourier Transforms
  - Includes the estimation of π_0 and of the bandwidth
  - Handles semi-supervised situations and truncated distributions
  - Produces graphics





kerfdr

□ Application 1: simulations

- p-values simulated according to the mixture model
- f_0 is the uniform distribution over [0,1]
- 4 proportions of null hypotheses: π_0 = 0.99 / 0.95 / 0.90 / 0.70
- f_1 is either an exponential $\epsilon(\mu_1)$ or a uniform distribution over $[0, 2\mu_1]$
- 2 different means for f_1: μ_1 = 0.01 / 0.001
- Number of observations: n = 1,000
- Number of simulations: S = 500

kerfdr

□ Application 1: simulations

- *p*-values simulated according to the mixture model
- f_0 is the uniform distribution over [0,1]
- 4 proportions of null hypotheses: π_0 = 0.99 / 0.95 / 0.90 / 0.70
- f_1 is either an exponential $\epsilon(\mu_1)$ or a uniform distribution over $[0,2\mu_1]$
- 2 different means for f_1: μ_1 = 0.01 / 0.001
- Number of observations: n = 1,000
- Number of simulations: S = 500
- Performances are assessed by means of the Root Mean Square Error :

$$RMSE(\pi_0, f) = \frac{1}{S}\sum_{s}\sqrt{\frac{1}{n}\sum_{i}\left(\widehat{\tau}_i^s - \tau_i\right)^2}.$$

expected value

estimated value

kerfdr

□ Application 1: simulations

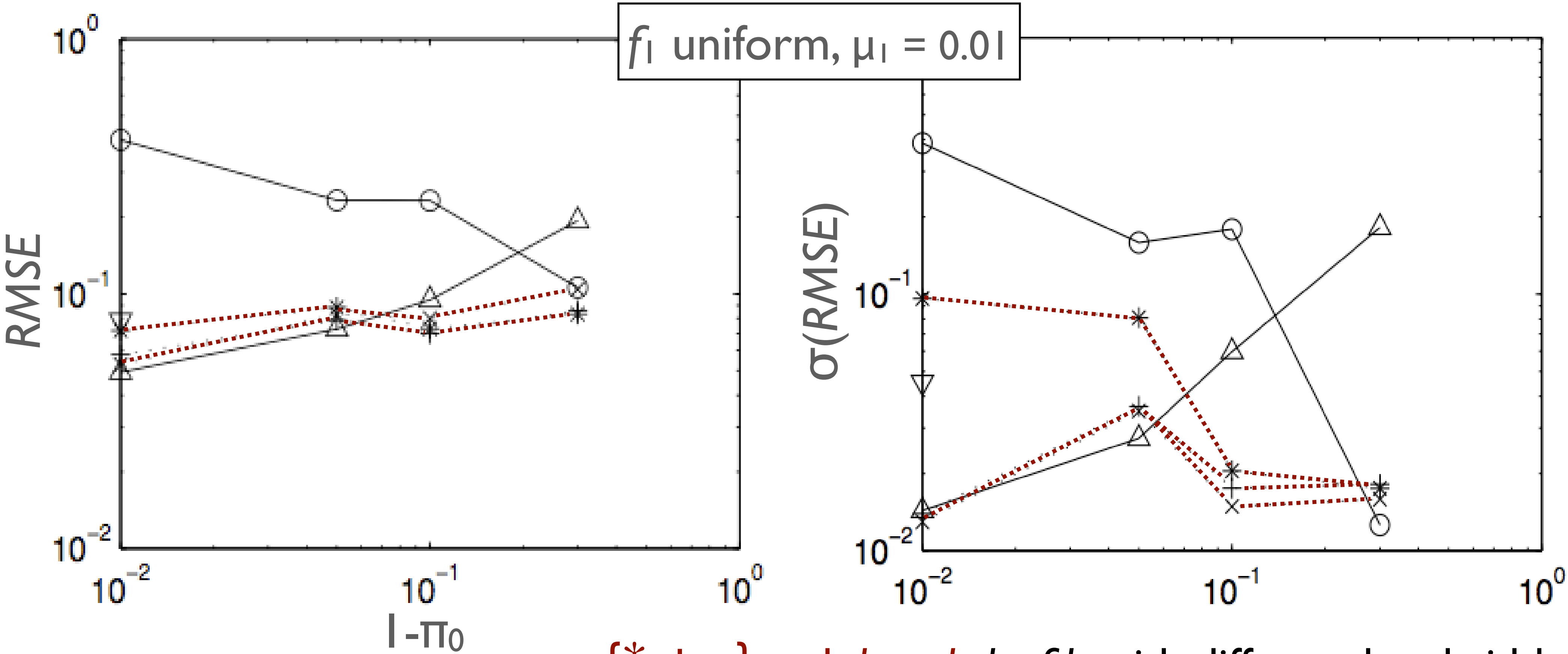
- *p*-values simulated according to the mixture model
- f_0 is the uniform distribution over [0,1]
- 4 proportions of null hypotheses: π_0 = 0.99 / 0.95 / 0.90 / 0.70
- f_1 is either an exponential $\epsilon(\mu_1)$ or a uniform distribution over $[0,2\mu_1]$
- 2 different means for f_1: μ_1 = 0.01 / 0.001
- Number of observations: n = 1,000
- Number of simulations: S = 500
- Performances are assessed by means of the Root Mean Square Error :

$$RMSE(\pi_0, f) = \frac{1}{S}\sum_s \sqrt{\frac{1}{n}\sum_i (\widehat{\tau}_i^s - \tau_i)^2}.$$

- **The smaller the *RMSE*, the better the performances.**

kerfdr

- □ Application 1: comparison with existing methods



f_1 uniform, $\mu_1 = 0.01$

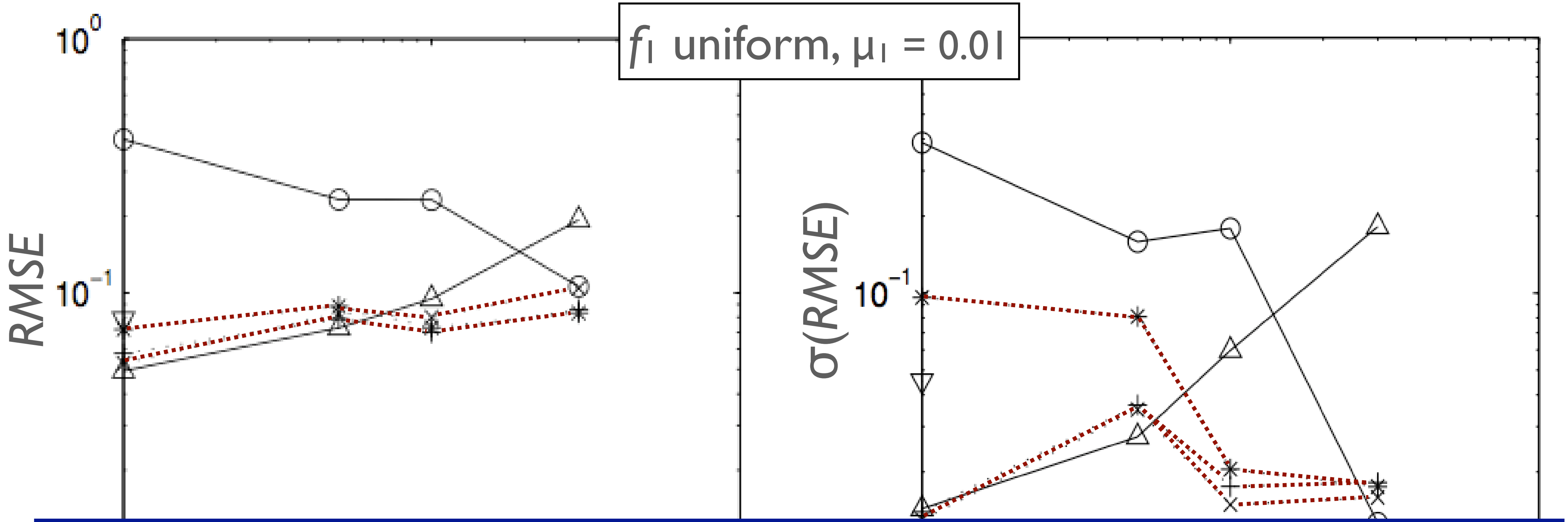
{*, +, *x*} and *dotted* : *kerfdr* with different bandwidth

-Δ- : Splines-based density estimation (Efron 04)

-O- : EM 2-components Gaussian mixture model (McLachlan et al 06)

kerfdr

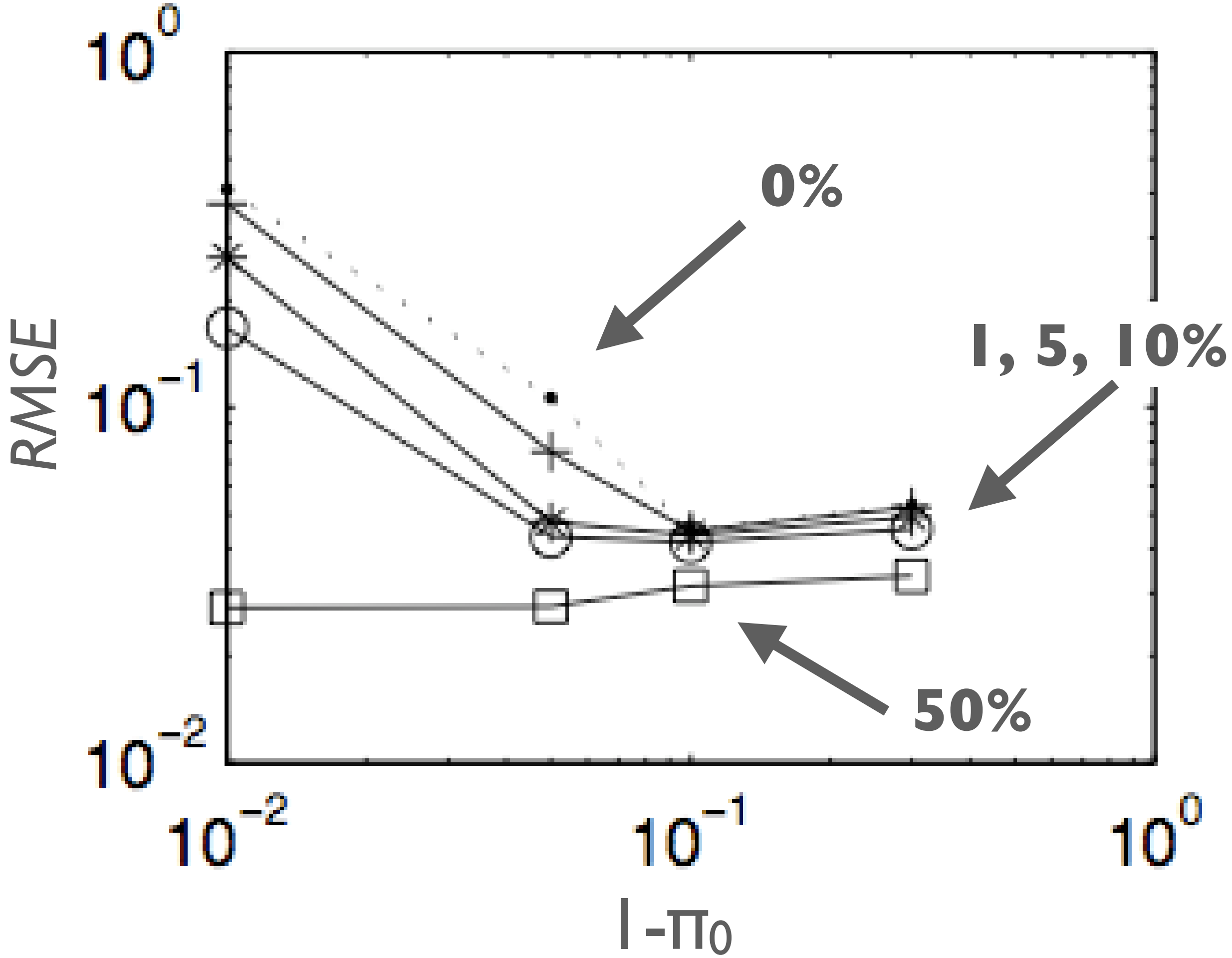
- Application 1: comparison with existing methods


f_1 uniform, μ_1 = 0.01

- Estimates of *kerfdr* not very sensitive to the bandwidth
- *kerfdr* performs as well the other methods when f_0 and f_1 are well separated (μ_1 = 0.001, data not shown)
- It outperforms them in more difficult situations (μ_1 = 0.01) especially in terms of stability.

kerfdr

- □ Application 1: semi-supervised : from 0% to 50% of known hypotheses



0%

1, 5, 10%

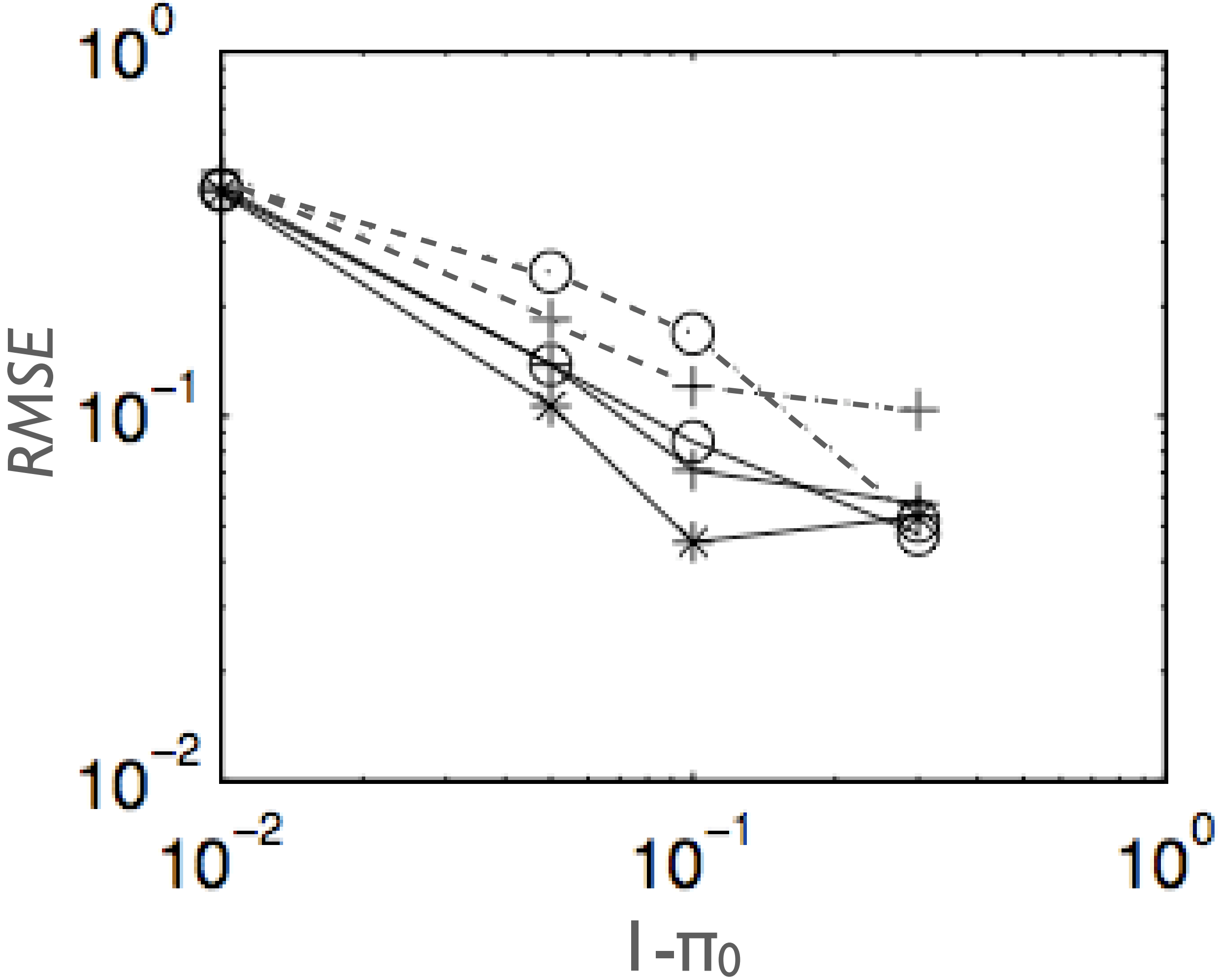
50%

The proportion of known hypotheses improves the estimates.

Even a small proportion of 1 or 5 % !!!

kerfdr

- □ Application 1: truncated distributions : p-value are truncated to a given threshold p^*



* : $p^* = 0$ (reference)

○ : $p^* = 10^{-3}$

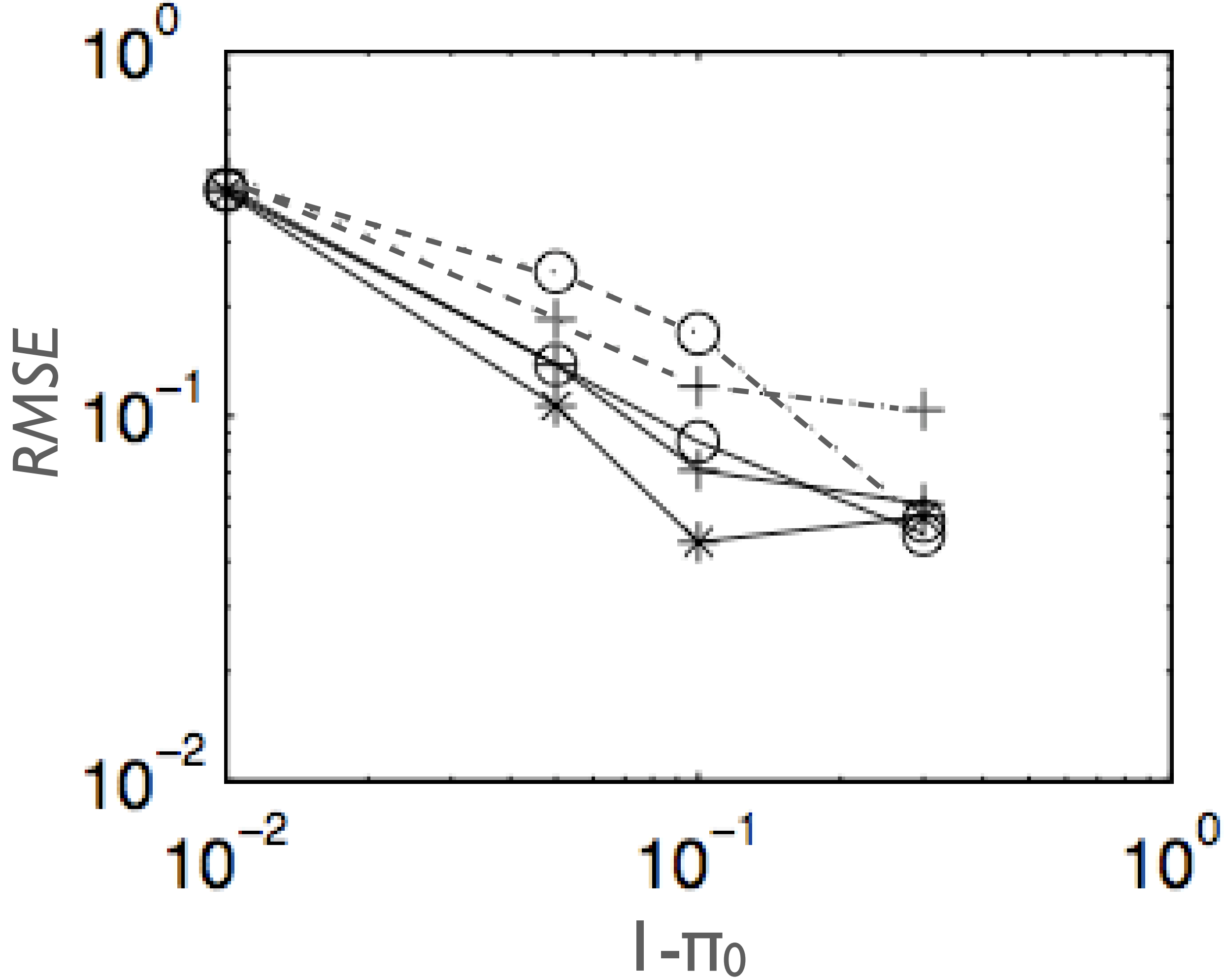
+ : $p^* = 10^{-2}$

dotted : naive estimation

lines : corrected estimation

kerfdr

- Application 1: truncated distributions : *p*-value are truncated to a given threshold *p**



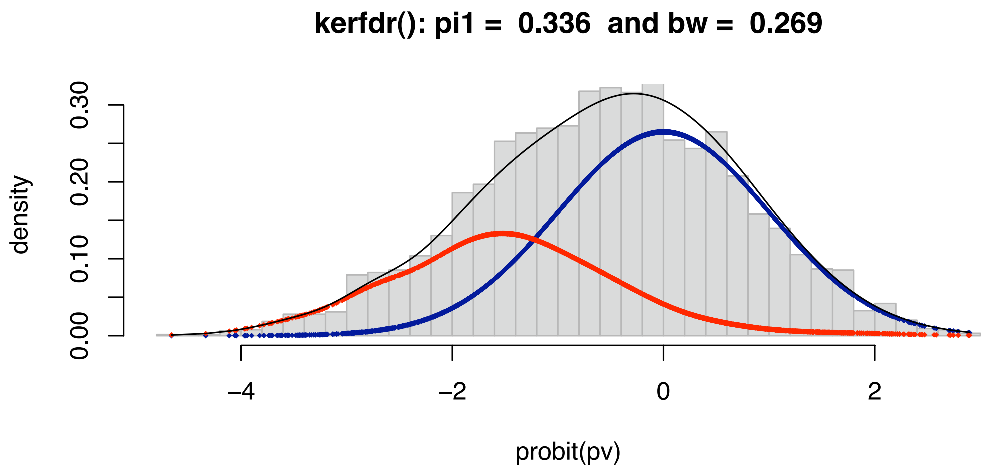
The correction improves the quality of the estimates.

The corrected estimates can be almost as good as the untrucated reference !!!

(Hedenfalk et al 01)

kerfdr

- **Application 2:** differential gene-expressions
  - 3,226 genes studied among two groups of BRCA1 (7 patients) and BRCA2 (8 patients).
  - Test: t test-like statistic (Delmar et al 05).

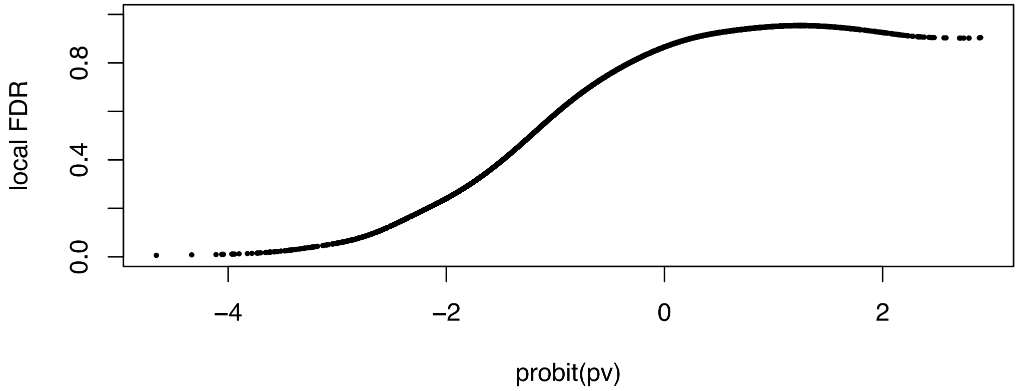


🟦 $f_0(x)$

🟥 $f_1(x)$

⬛ $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$

- 1 - π_0 = 0.336
- # of genes < 1% = 5
- running time < 1 sec



(data provided by Merck-Serono)

kerfdr

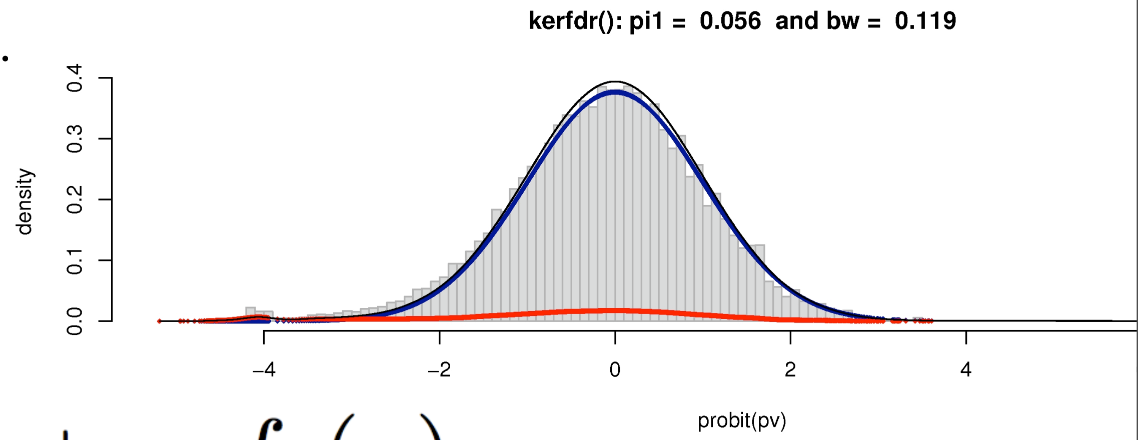
□ Application 3: genome-wide association

- □ 203 controls from Rennes genotyped using a 100K Affy (100,000 SNPs covering the genome).
- □ Test: Hardy-Weinberg equilibrium test.

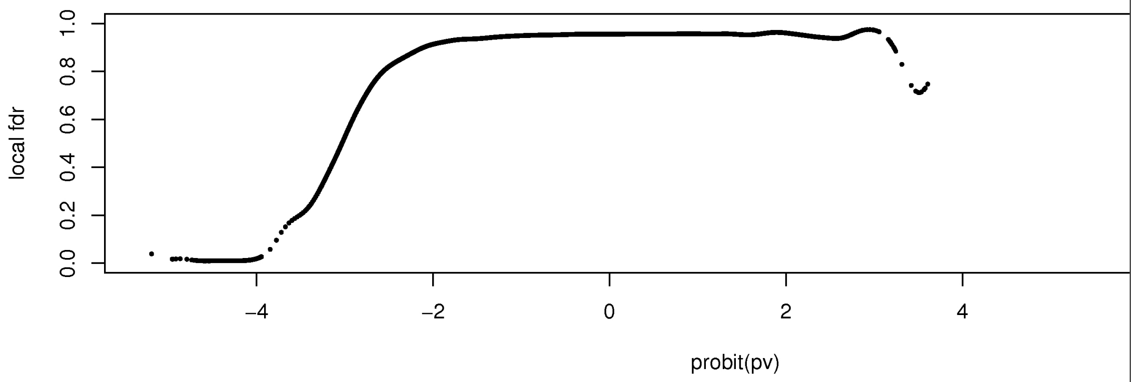
🟦 $f_0(x)$

🟥 $f_1(x)$

⬛ $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$



- $1 - \pi_0 = 0.056$
- # of SNPs < 1% = 29
- running time < 3 sec



kerfdr

□ Initial method fully described in *Robin et al* 07.

□ Algorithm available *via* the CRAN or at

http://stat.genopole.cnrs.fr/software/kerfdr

□ Manuscript under revision in BMC Bioinformatics.

Acknowledgements

the Statistics for System Biology working group

the Statistics and Genome laboratory, Evry, FRANCE

Merck-Serono for the data.

S Robin, A Bar-Hen and JJ Daudin for the initial method

e-mail: mickael.guedj@gmail.com

Statistique et Génome

UNIVERSITÉ D'EVRY
VAL D'ESSONNE

serono
biotech & beyond

SSB for Statistics Systems Biology

Any questions ??

« That's what I want to say. See if you can find some statistics to prove it! »