Differential analysis of microarray data, Multiple testing problems and Local False Discovery Rate.

S. Robin
robin@inapg.inra.fr

UMR INA-PG / INRA, Paris
Mathématique et Informatique Appliquées

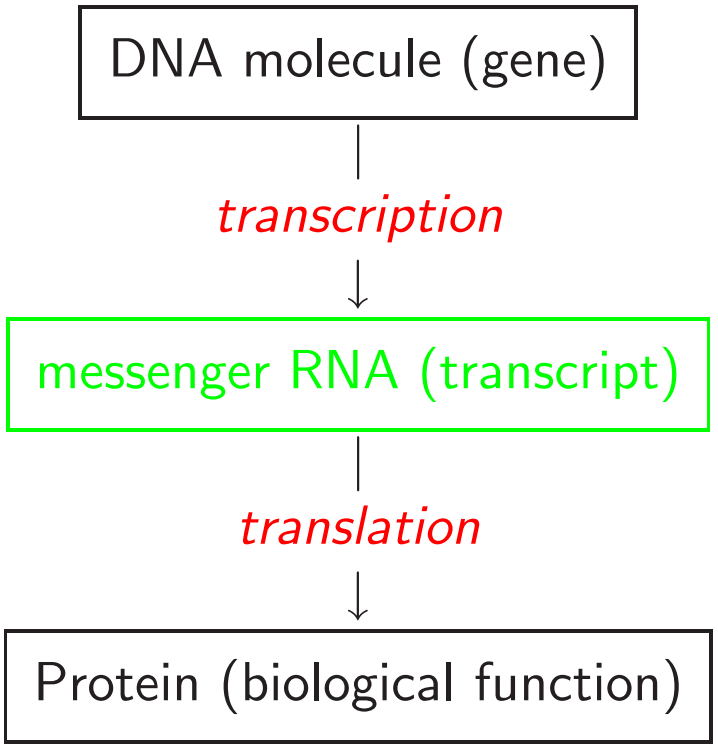
Semi-parametric modeling

joint work with J.-J. Daudin, A. Bar-Hen, L. Pierre

Bio-Info-Math Workshop, Tehran, April 2005

Microarray data and differential analysis

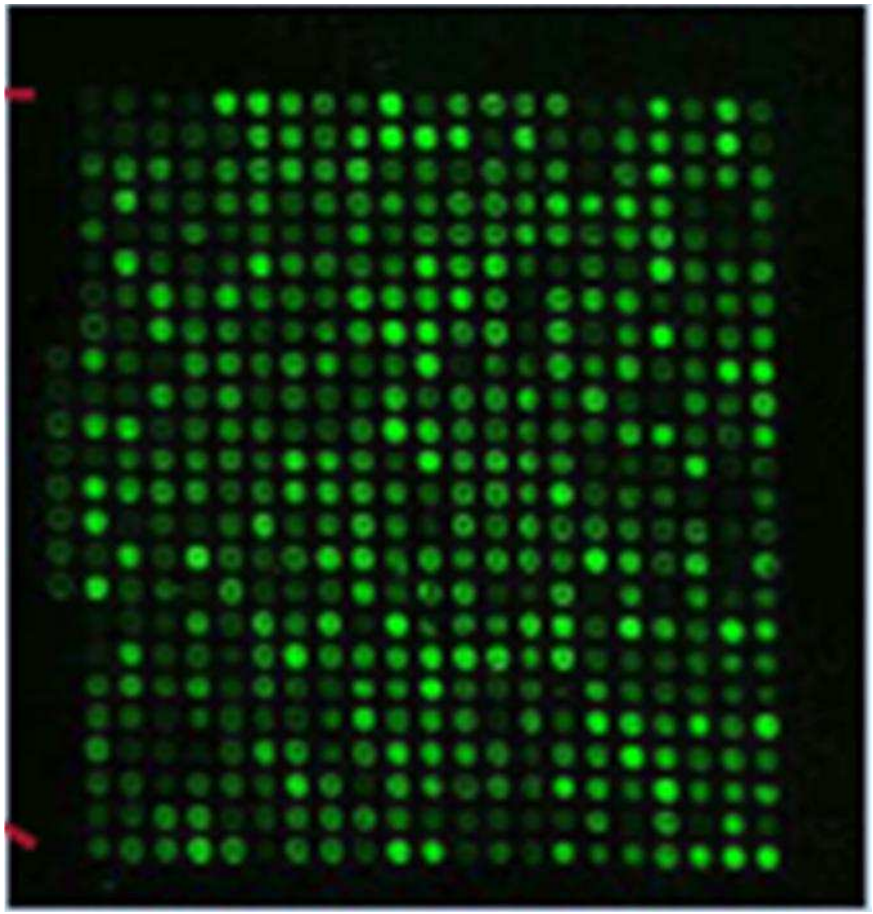
Molecular biology central dogma



"Definition": $\left(\begin{array}{c}\text{Expression level}\\ \text{of a gene}\end{array}\right) \propto \left(\begin{array}{c}\text{number of copies}\\ \text{of mRNA in the cell}\end{array}\right)$

Microarray technology

Aims to monitor the expression level of several thousands of genes simultaneously

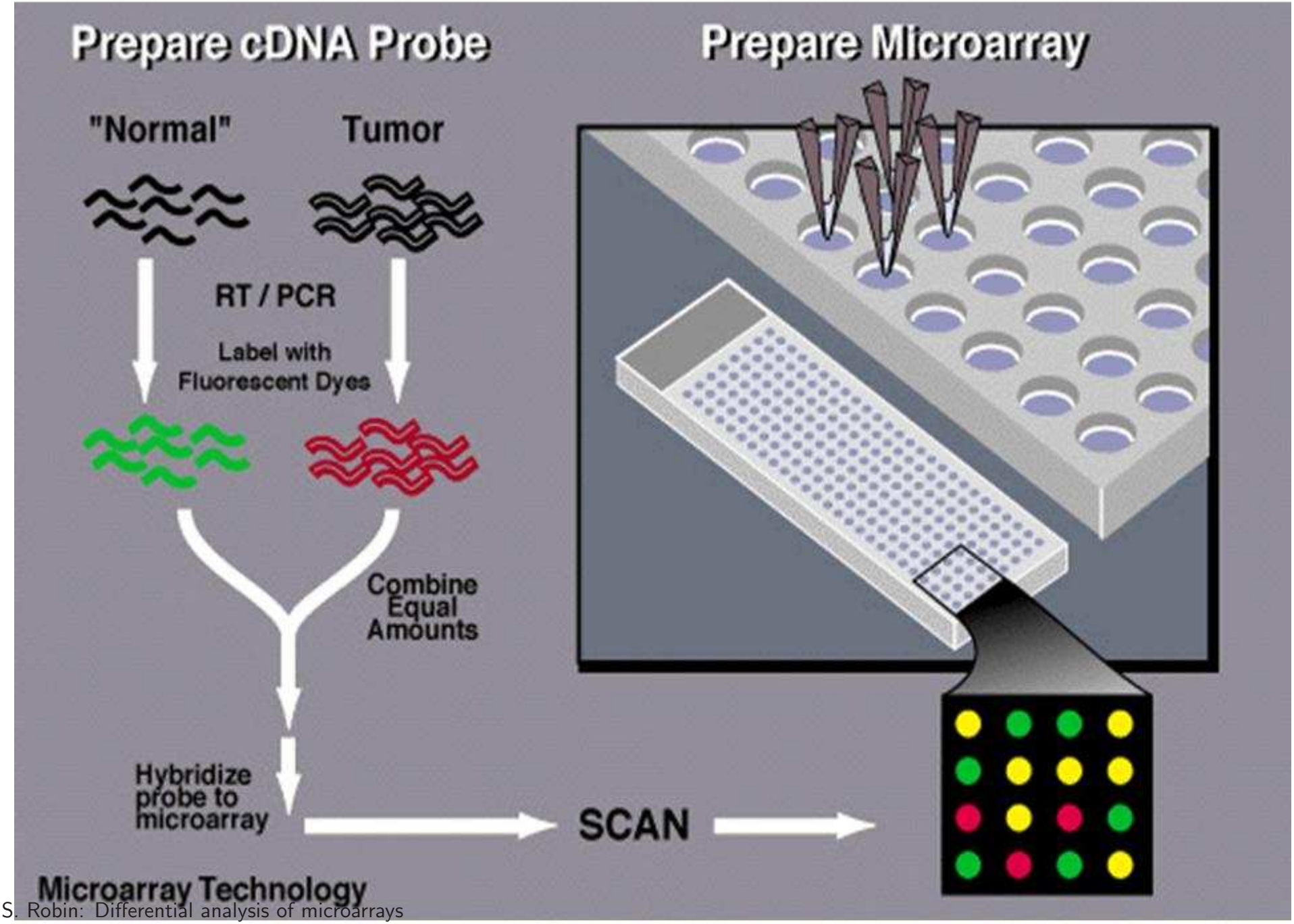


1 spot = 1 gene

Expression level in the cell:

- at given time,
- in a given condition

Inferring genes' functions. Determining the conditions (times, tissues, $etc.$) in which the expression of a given gene is the highest (or lowest) should help in understanding its function.


Prepare cDNA Probe
Prepare Microarray
"Normal"
Tumor
RT / PCR
Label with
Fluorescent Dyes
Combine
Equal
Amounts
Hybridize
probe to
microarray
SCAN
Microarray Technology

Differential analysis

Elementary data: Y_{itr} = expression level of gene i in condition t ($t = 1$ or 2) at replicate r

Differentially expressed genes are genes for which Y_{i1r} is not distributed as Y_{i2r}.

Null hypothesis for gene i: $\mathbf{H}_0(i) = \{Y_{i1r} \stackrel{\mathcal{L}}{=} Y_{i2r}\}$

Statistical test: Student, Wilcoxon, permutation, $etc.$

For each gene we get:

the value of the test statistic T_i

the corresponding p-value $P_i = \Pr\{\mathcal{T} > T_i \mid \mathbf{H}_0(i)\}$

Comparing more than 2 conditions. Same problem: Fisher, Kruskall-Wallis tests provide one p-value for each gene.

Multiple testing problem

Rejection rule: For a given level α,

$$P_i < \alpha \qquad \Longrightarrow \qquad \text{gene } i \text{ is declared positive (i.e. differentially expressed)}$$

Multiple testing: When performing n simultaneous tests

	Decision (random)		
	$\mathbf{H}_0$ accepted	$\mathbf{H}_0$ rejected	
$\mathbf{H}_0$ true	TN true negatives	FN false negatives	n_0 negatives
$\mathbf{H}_0$ false	FP false positives	TP true positives	n_1 positives
	N negatives	R positives	n

All the random quantities (capital) depend on the data and the pre-fixed level α.

Microarray experiment: Typically $n = 10\,000$ tests are performed simultaneously

For $\alpha = 5\%$, if no gene is actually differentially expressed ($n_1 = 0, n_0 = n$), we expect

$$0.05 \times 10\,000 = 500 \text{ "positive" genes}$$

which are all false positives.

Problem: We'd like to control some "global risk" α^* such as

- the probability of having one false positive (FWER)

$$FWER = \Pr\{FP \geq 1\},$$

- or the proportion of false positives (FDR)

$$FDR = \mathbb{E}(FP/R).$$

(Benjamini & Hochberg, JRSS-B, 1995; Dudoit & al., Stat. Sci., 2003)

Family Wise Error Rate (FWER)

$$FWER = \Pr\{FP \geq 1\}$$

Sidak: If the n tests are independent,

$$FP \sim \mathcal{B}(n, \alpha) \qquad \Longrightarrow \qquad \Pr\{FP \geq 1\} = 1 - (1-\alpha)^n.$$

Fixing level at $\alpha = 1 - (1-\alpha^*)^{1/n} (\simeq \alpha^*/n)$ ensures $FWER = \alpha^*$.

Bonferroni: In any case

$$FWER = \Pr\left\{\bigcup_i i \text{ false positive}\right\} \leq \sum_i \Pr\left\{i \text{ false positive}\right\} = n\alpha$$

Fixing level at $\alpha = \alpha^*/n$ ensures $FWER \leq \alpha^*$.

Remark: The independent case is, in some sense, the worst case.

Adaptive procedure for FWER

Idea:

One step procedure are designed for the smallest p-value

$\Longrightarrow$ they are too conservative.

Principle: Order the p-values

$$P_{(1)} \leq \cdots \leq P_{(i)} \leq \cdots < P_{(n)}.$$

Step 1: Apply (say) the Bonferroni correction to $P_{(1)}$: if $P_{(1)} \leq \alpha^*/n$ then go to step 2

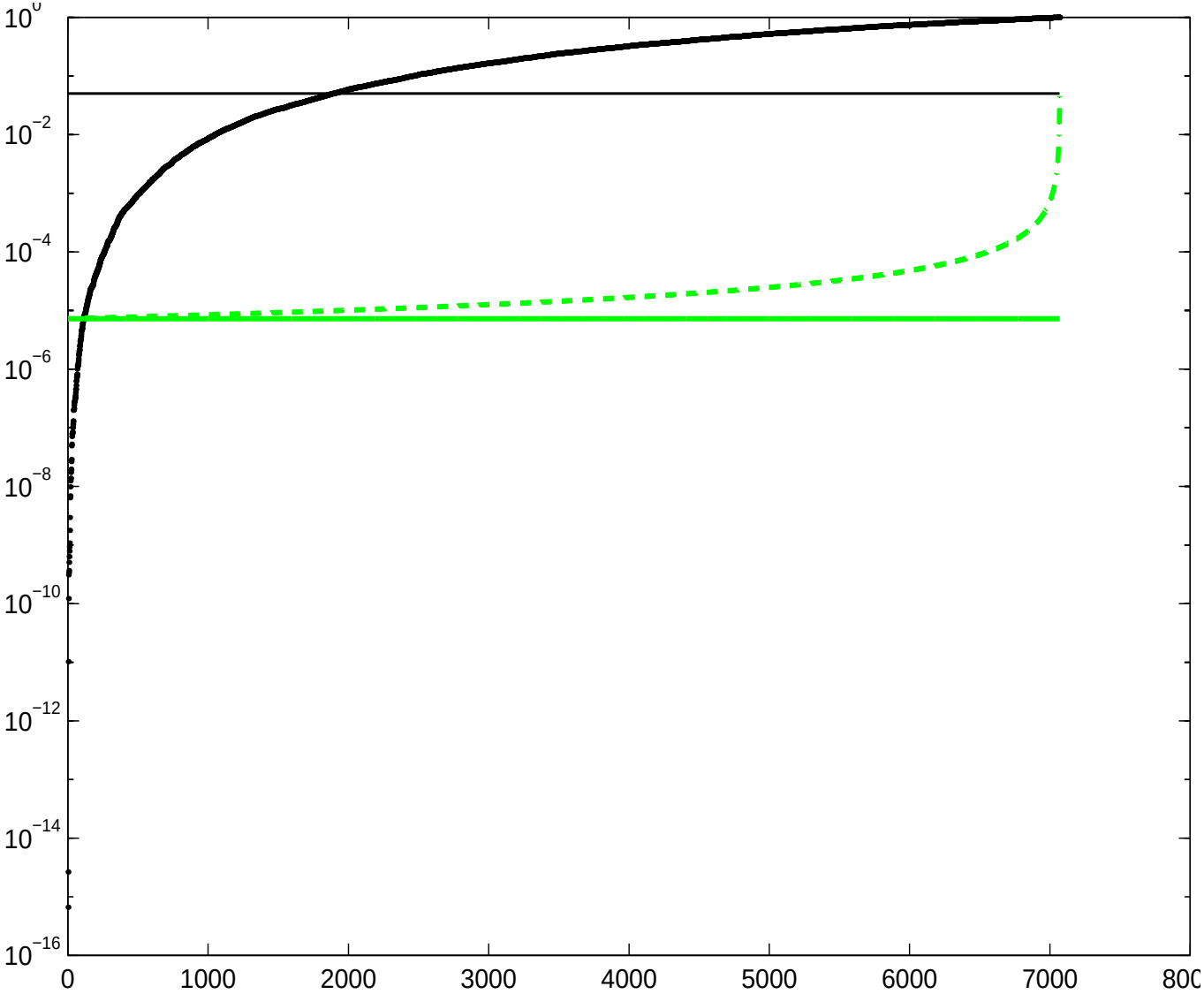
Step 2: Apply the same correction to $P_{(2)}$, replacing n by $n-1$:

$$\text{if } P_{(2)} \leq \alpha^*/(n-1) \text{ then go to step 3}$$

Step k: Apply the same correction to $P_{(k)}$, replacing n by $n-k+1$:

$$\text{if } P_{(k)} \leq \alpha^*/(n-k+1) \text{ then go to step } k+1$$

Thresholds for Golub data: 27 patients with AML, 11 with ALL, $n = 7070$ genes, Welch test



. . . p-value

– 5%

– Bonferroni

. . . Holm

– Sidak

. . . Sidak ad.

Adjusted p-values can be directly compared to the desired FWER α^*.

- One step Bonferroni

$$P_{(i)} \leq \alpha^*/n \qquad \Longleftrightarrow \qquad \tilde{P}_{(i)} = \min(nP_{(i)}, 1) \leq \alpha^*$$

- One step Sidak

$$P_{(i)} \leq 1 - (1-\alpha^*)^{1/n} \qquad \Longleftrightarrow \qquad \tilde{P}_{(i)} = 1 - (1-P_{(i)})^n \leq \alpha^*$$

- Adaptive Bonferroni (Holm, 79)

$$\tilde{P}_{(i)} = \max_{j \leq i} \{\min[(n-j+1)P_{(j)}, 1]\}$$

- Adaptive Sidak

$$\tilde{P}_{(i)} = \max_{j \leq i} \{\min[1 - (1-P_{(j)})^{n-j+1}, 1]\}$$

Accounting for dependency

The Westfall & Young (93) procedure preserves the correlation between genes using permutation tests and applying the same permutations to all the genes.

Adjusted p-values are estimated by

$$\hat{\tilde{p}} = \frac{1}{S}\sum_s \mathbb{I}\{p^s_{(g)} < p_g\} \qquad \text{"minP" procedure}$$

$$\frac{1}{S}\sum_s \mathbb{I}\{|T^s_{(g)}| > |T_g|\} \qquad \text{"maxT" procedure}$$

The same procedure allows to estimate the distribution of the second, third, etc., smallest p value

Limitation. The number of replicates strongly conditions the precision of the estimated distribution:

$$\left(\begin{array}{c} 8 \\ 4 \end{array}\right) = 70, \qquad \left(\begin{array}{c} 10 \\ 5 \end{array}\right) = 252$$

False Discovery Rate (FDR)

$$FDR = \mathbb{E}(FP/R)$$

Idea: Instead of preventing any error, just control the proportion of errors
$\Longrightarrow$ less conservative

Benjamini & Hochberg (95) procedure: Given the sorted p-values

$$P_{(1)} \leq \cdots \leq P_{(i)} \leq \cdots \leq P_{(n)},$$

rejecting $\mathbf{H}_0$ for all (i) such as

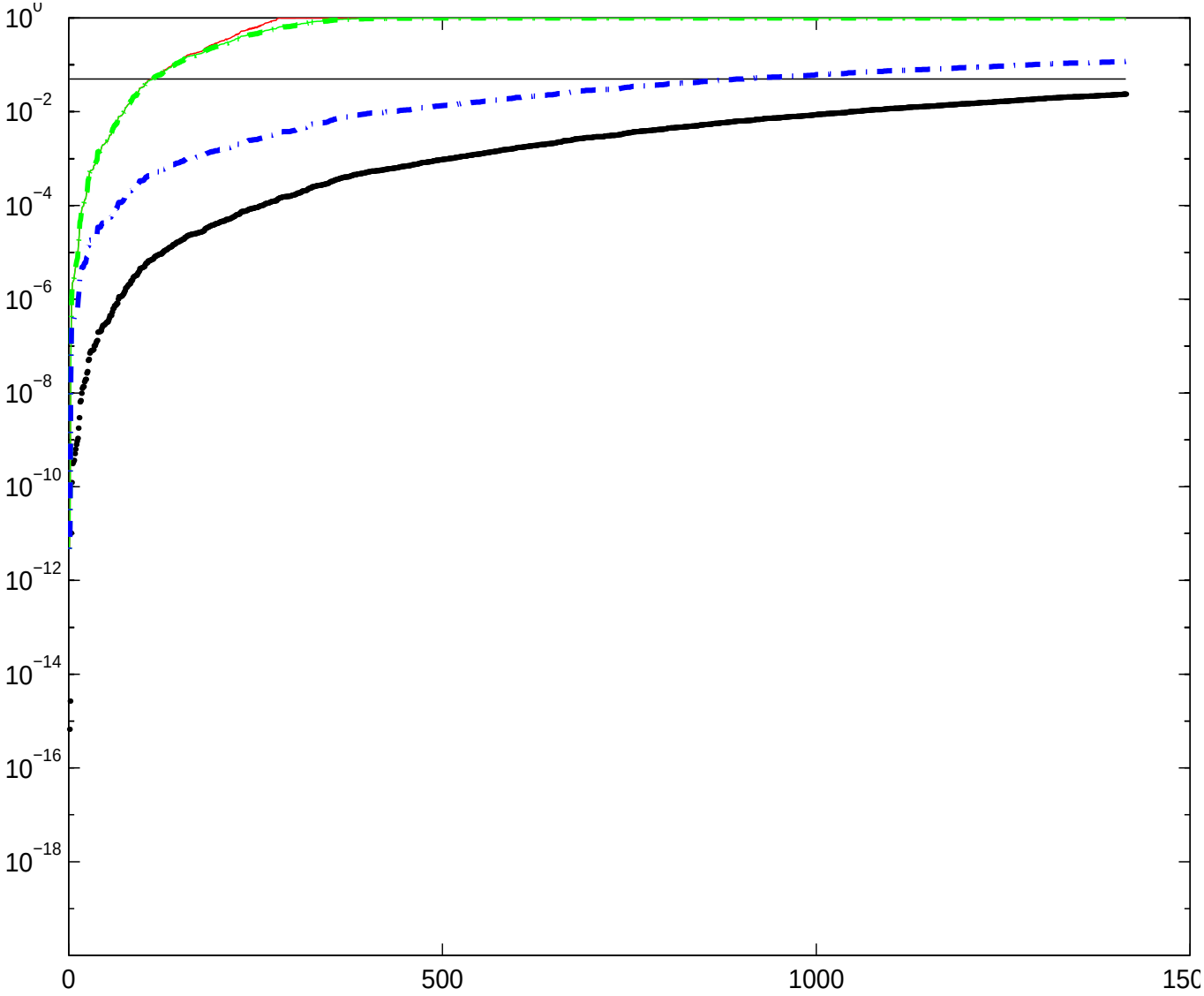
$$P_{(i)} \left(\leq \frac{i\alpha^*}{n} \right) \leq \frac{i\alpha^*}{n_0} \qquad \Longrightarrow \qquad FDR \leq \frac{n_0}{n}\alpha^* \leq \alpha^*$$

Benjamini & Yakutieli (01): For positively correlated test statistics

$$P_{(i)} \leq \frac{i\alpha^*}{n(\sum_j 1/j)}.$$

Adjusted p-values for Golub data / Number of positive genes: $\alpha^* = 5\%$

p-value:	1887
Bonferroni:	111
Sidak:	113
Holm:	112
Sidak adp.:	113
FDR:	903



Local False Discovery Rate

FDR provides a general information about the risk of the whole procedure (up to step i).

We are interested in a specific risk, associated to each gene.

Local FDR (ℓFDR). First defined by Efron & al. (JASA, 2001) in a mixture model framework:

$$\ell FDR_i := \Pr\{\mathbf{H}_0(i) \text{ is false} \mid T_i\}.$$

Derivative of the FDR: $\ell FDR_{(i)}$ can be also defined as the derivative of the FDR

$$\ell FDR(t) = \lim_{h \downarrow 0} \frac{FDR(t+h) - FDR(t)}{h}$$

which can be estimated by

$$\widehat{n}_0 (P_{(i)} - P_{(i-1)})$$

(Aubert & al., BMC Bioinfo., 04).

Estimation of the proportion n_1/n

The efficiency of all multiple testing procedures would be improved if n_0 was known.

Empirical cdf. The cumulative distribution function (cdf) of the p-value can be estimated via its empirical version:

$$\widehat{G}(p) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{P_i \leq p\}.$$

The cdf of the negative p-values is given by the uniform distribution:

$$\Pr\{P_i \leq p \mid i \in \mathbf{H}_0\} = p.$$

Cdf mixture. Denoting F the cdf of the positive p-value, we have

$$G(p) = aF(p) + (1-a)p, \qquad \text{where} \qquad a = n_1/n.$$

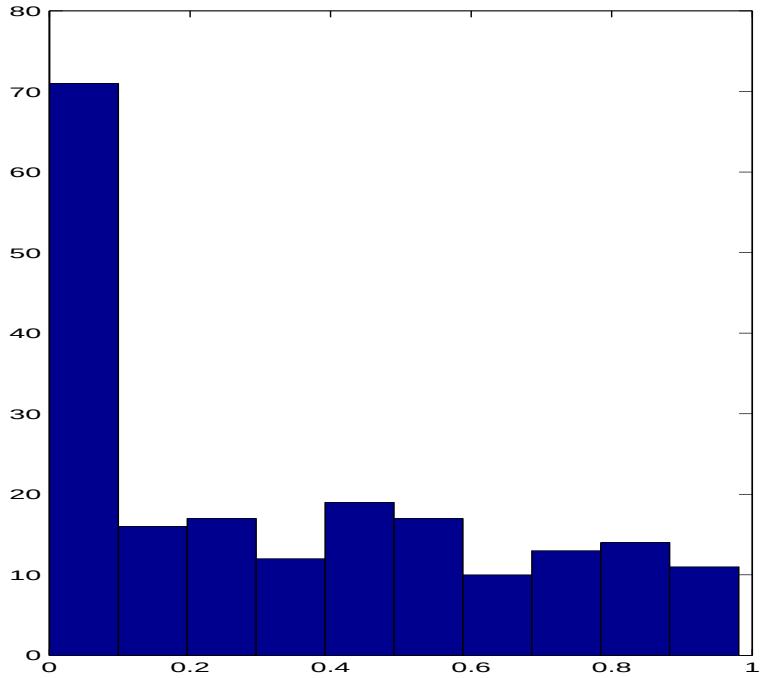
Above a certain threshold t, $F(p)$ should be close to 1:

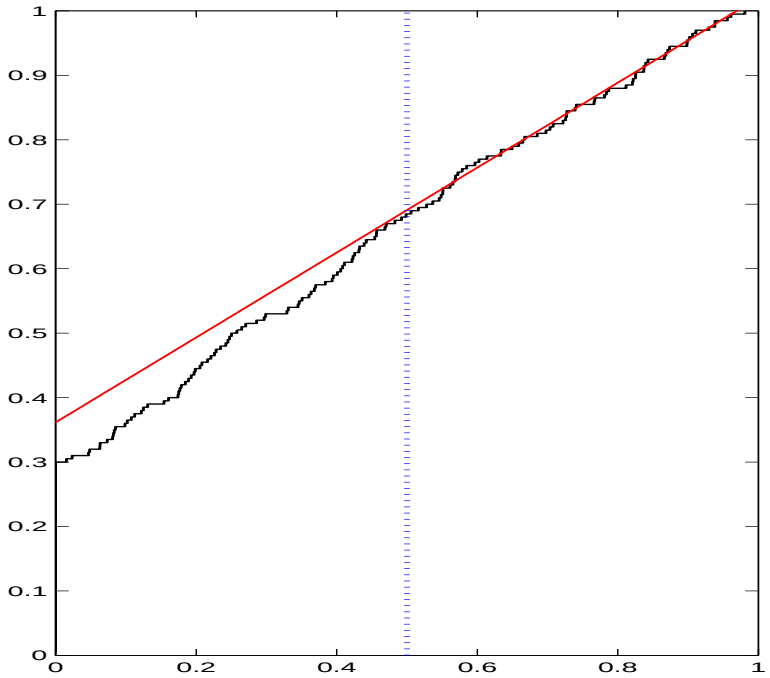
$$x > t: \quad G(p) \simeq a + (1-a)p.$$

Empirical proportion. Storey & al, Genovese & Wasserman (JRSS-B, 02) propose an estimate of a based on this approximation:

$$\widehat{a} = [1 - P(t)/n]/(1 - t).$$

Linear regression. $(1 - a)$ can also be estimated by the coefficient of the linear regression of $\widehat{G}(p)$ wrt p





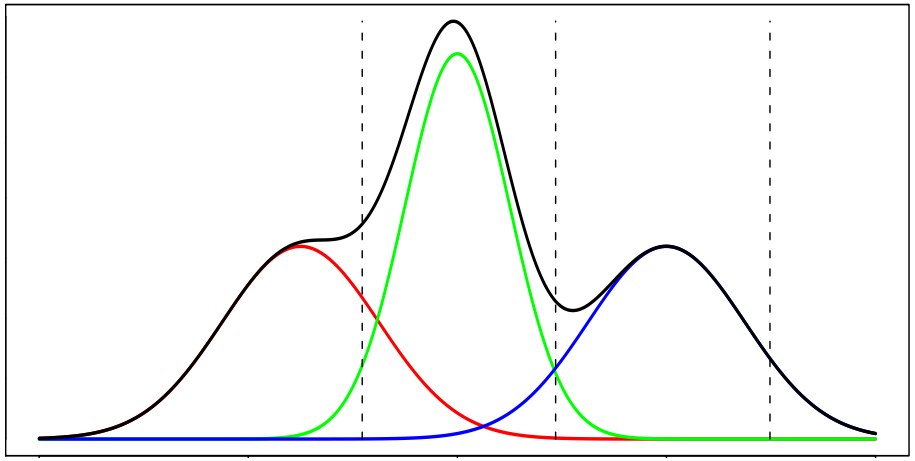
Mixture model

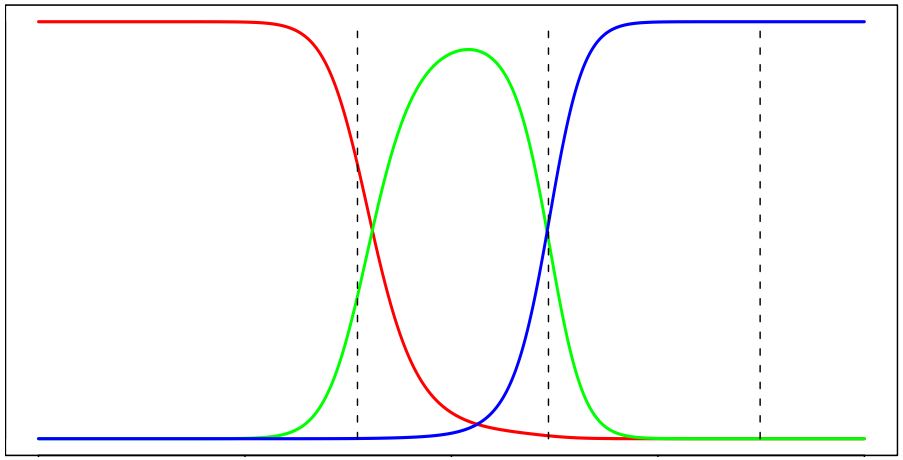
Model:

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x)$$

Posteriori probability:

$$\tau_{gk} = \Pr\{g \in f_k \mid x_g\} = \pi_k f_k(x_g)/f(x_g)$$



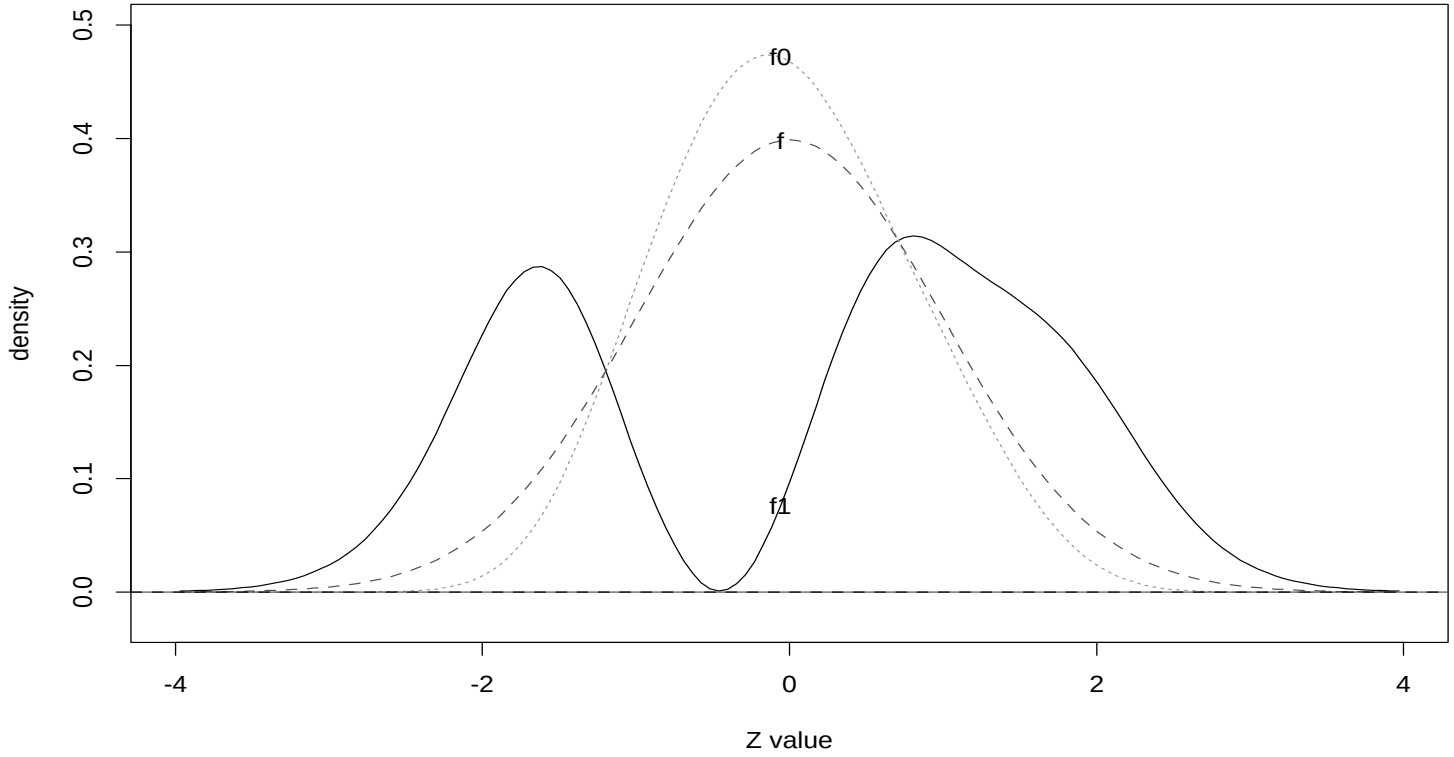


τ_{gk} (%)	$g=1$	$g=2$	$g=3$
$k=1$	65.8	0.7	0.0
$k=2$	34.2	47.8	0.0
$k=3$	0.0	51.5	1.0

Distribution of the test statistic. Efron & al. (01) propose to describe the distribution of the test statistic T_i using a mixture model.

$$T_i \sim f(t) = p_1 f_1(t) + p_0 f_0(t)$$

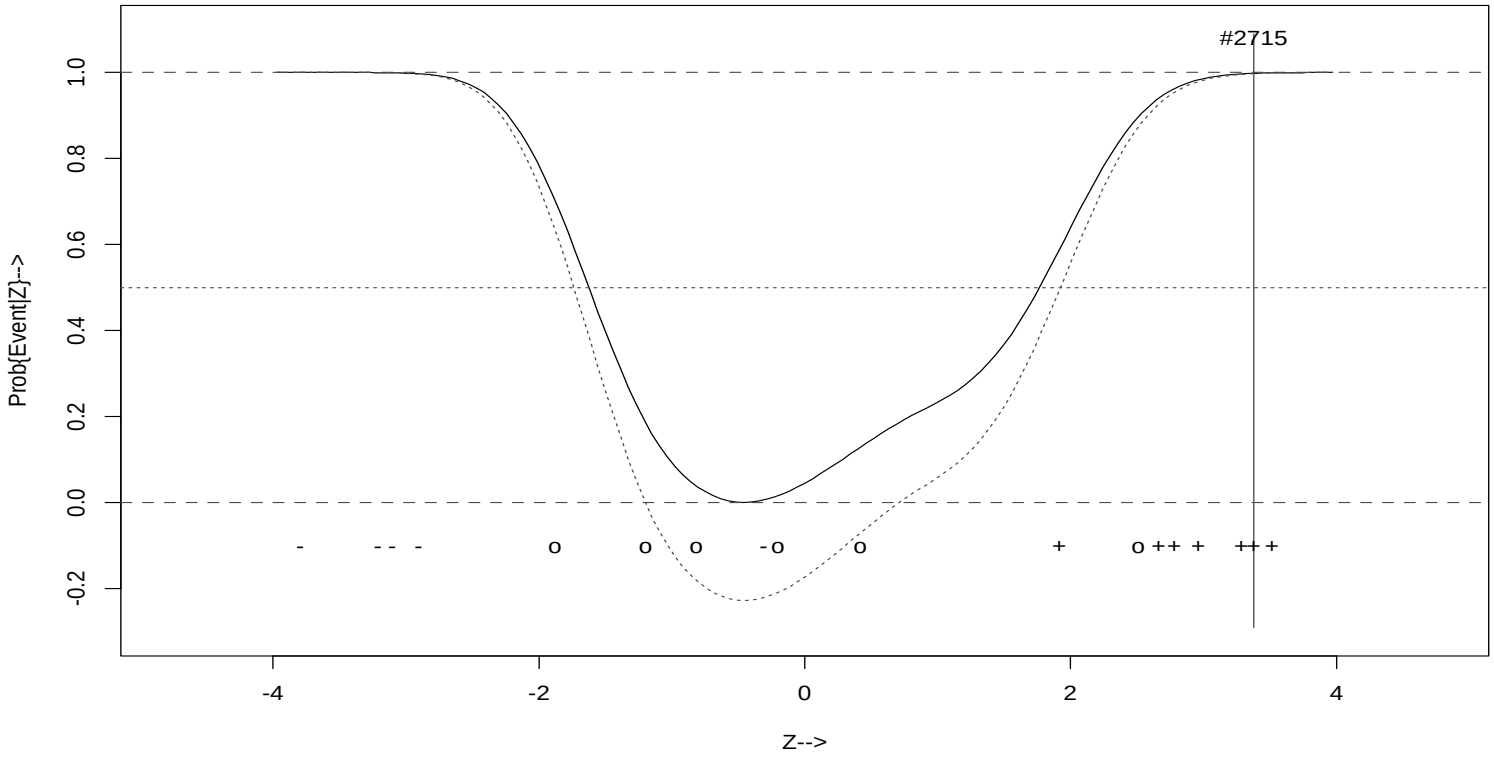
where both, a, f_0 and f_1 have are unknown.



Using the high number of replicates of their example (Affymetrix data), they use a local logistic regression to estimate the *local FDR* (ℓFDR):

$$\ell FDR_i = p_0 f_0(T_i)/f(T_i)$$

which is actually the *posterior probability* that the test i is actually negative given the value of the test statistic.



Mixture for the p-values

Allison (02) proposes the same strategy regarding the p-values, assuming that

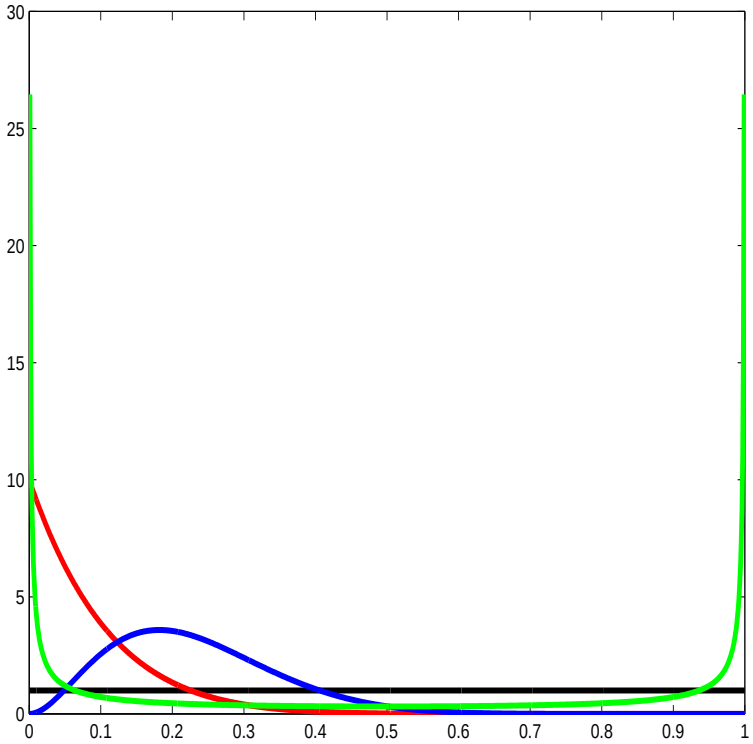
$$P \sim aB(r,s) + (1-a)\mathcal{U}_{[0;1]}$$

where the proportion a and the parameters r and s have to be estimated, for example, using the E-M algorithm.

Beta density:

$$\beta(p;r,s) = \frac{p^{r-1}(1-p)^{s-1}}{B(r,s)},$$

$$0 \leq p \leq 1.$$



E-M algorithm. The most popular algorithm to estimate the parameters of a mixture model is Expectation-Maximization. The principle is to alternate the two steps.

E step: For each observation i calculate the posterior probability τ_i that it comes from the non-null distribution using Bayes' formula

$$\tau_i^{h+1} = \frac{\widehat{a}^h \beta(p_i; \widehat{s}^h, \widehat{r}^h)}{\widehat{g}^h(p_i)}, \qquad \widehat{g}^h(p_i) = \widehat{a}^h \beta(p_i; \widehat{r}^h, \widehat{s}^h) + (1 - \widehat{a}^h)$$

M step: Calculate the maximum-likelihood estimates of r and s giving to each observation i a weight τ_i^{h+1}.

Properties:

1. At each E-M step, the likelihood of the data under the mixture model increases.

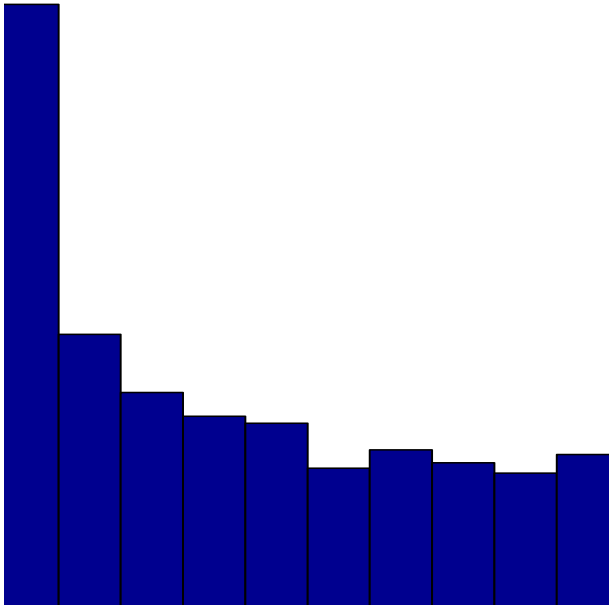
2. E-M provide estimates of the posterior probabilities which are actually the most relevant quantities.

Semi-parametric mixture model

Mixture model

Property of the test statistic. The standard hypotheses testing theory implies that, under $\mathbf{H}_0(i)$, P_i is uniformly distributed over $[0,1]$:

$$P_i \underset{\mathbf{H}_0(i)}{\sim} \mathcal{U}_{[0,1]}$$



The P_i's are distributed according to a mixture distribution with density

$$g(p) = af(p) + (1-a)$$

The problem is then to estimate

- a: the proportion of differentially expressed genes
- f: the (alternative) density f

Generalization: We consider an i.i.d. sample $\{X_1, \ldots, X_n\}$ with mixture density

$$g(x) = af(x) + (1-a)\phi(x)$$

The proportion a is unknown $\longrightarrow$ parametric part

The density f is completely unknown $\longrightarrow$ non parametric part

The density ϕ in completely specified ($\mathcal{U}_{[0,1]}$, $\mathcal{N}(0,1)$, $etc.$)

Posterior probability. We are interested in the estimation of

$$\tau_i = \Pr\{Z_i = 1 \mid x_i\} = \mathbb{E}(Z_i \mid x_i) = \frac{af(x_i)}{g(x_i)}$$

where $Z_i = \begin{cases} Z_i = 1 & \text{if } i \text{ comes from } f \quad (\mathbf{H}_0(i) \text{ false}), \\ Z_i = 0 & \text{otherwise} \qquad\qquad (\mathbf{H}_0(i) \text{ true}). \end{cases}$

Density estimation

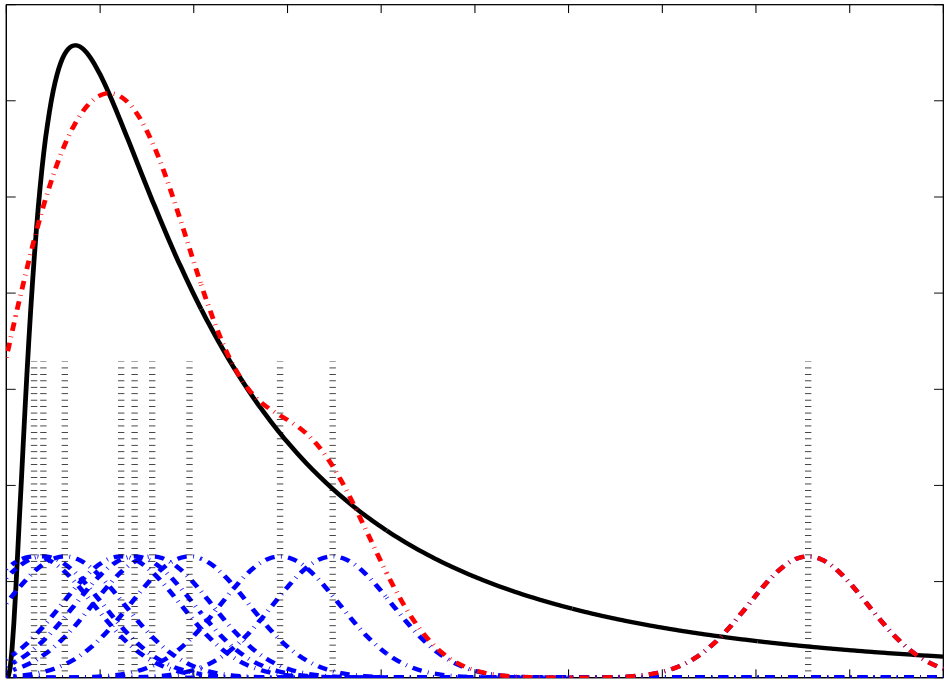
Kernel estimate. A natural non-parametric estimate of f is

$$\widehat{f}(x) = \frac{1}{\sum_i Z_i} \sum_i Z_i k_i(x)$$

where

$$k_i(x) = \frac{1}{h} k\left(\frac{x - x_i}{h}\right)$$

k being a kernel, i.e. a symmetric density function with mean 0.



Weighted kernel estimate. Since the Z_i's are unknown, we propose to replace them by their conditional expectations:

$$\widehat{f}(x) = \frac{1}{\sum_i \tau_i} \sum_i \tau_i k_i(x)$$

τ_i is the weight of observation i in the estimation of f.

Property of the $\widehat{\tau}_i$. The estimates of the τ_i's must satisfy

$$\widehat{\tau}_j = \frac{a\widehat{f}(x_j)}{\widehat{g}(x_j)} = \frac{a\sum_i \widehat{\tau}_i k_i(x_j)}{a\sum_i \widehat{\tau}_i k_i(x_j) + (1-a)\phi(x_j)\sum_i \widehat{\tau}_i}$$

or

$$\widehat{\tau}_j = \frac{\sum_i \widehat{\tau}_i b_{ij}}{\sum_i \widehat{\tau}_i b_{ij} + \sum_i \widehat{\tau}_i} \qquad \text{with} \qquad b_{ij} = \frac{a}{1-a}\frac{k_i(x_j)}{\phi(x_j)} \geq 0$$

Function $\boldsymbol{\psi}$.

$$\begin{aligned} \boldsymbol{\psi} : \mathbb{R}^n &\rightarrow \mathbb{R}^n \\ \mathbf{u} &\rightarrow \boldsymbol{\psi}(\mathbf{u}) : \psi_j(\mathbf{u}) = \frac{\sum_i u_i b_{ij}}{\sum_i u_i b_{ij} + \sum_i u_i} \end{aligned}$$

$\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_n)$ is a fixed point of $\boldsymbol{\psi}$.

Estimation algorithm of $\boldsymbol{\tau}$. Given some initial $\widehat{\boldsymbol{\tau}}^0$, iterate $\boldsymbol{\psi}$:

$$\widehat{\boldsymbol{\tau}}^{h+1} = \boldsymbol{\psi}(\widehat{\boldsymbol{\tau}}^h).$$

a remains fix: it has to be estimated independently.

2 steps of the algorithm:
"E" step: given $\widehat{f}^h$ and $\widehat{g}^h$, calculate

$$\widehat{\boldsymbol{\tau}}^{h+1} = a\widehat{f}^h(x_i) \,/\widehat{g}^h(x_i)\,.$$

Other step: given $\widehat{\boldsymbol{\tau}}^h$, estimate f and g:

$$\widehat{f}^h(x) = \sum_i \widehat{\tau}_i^h k_i(x) \Bigg/ \sum_i \widehat{\tau}_i^h\,, \qquad \widehat{g}^h(x) = a\widehat{f}^h(x) + (1-a)\phi(x).$$

This second step does not maximize the likelihood $\longrightarrow$ not an E-M algorithm.

Theorem: $\boldsymbol{\psi}$ is contracting

$\Longrightarrow$ the algorithm converges toward its unique fix point.

Sketch of proof. $\boldsymbol{\psi} = \boldsymbol{\alpha} \circ \boldsymbol{\beta} \circ \boldsymbol{\gamma}$:

$$\alpha_j(\mathbf{u}) = \frac{u_j}{u_j + 1}, \qquad \beta_j(\mathbf{u}) = \sum_i b_{ij} u_i, \qquad \gamma_j(\mathbf{u}) = \frac{u_j}{\sum_i u_i},$$

1. Simplex $\mathcal{E} = \{\mathbf{u} : \sum_i u_i = 1\}$ ($\boldsymbol{\gamma}$ = projection on $\mathcal{E}$)

$$\mathbf{u}^* \in \mathbb{R}^n : \boldsymbol{\psi}(\mathbf{u}) = \mathbf{u} \qquad \Longleftrightarrow \qquad \mathbf{v}^* = \boldsymbol{\gamma}(\mathbf{u}^*) \in \mathcal{E} : \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v}) = \mathbf{v}$$

$\rightarrow$ Just consider $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ on the simplex $\mathcal{E}$.

2. Brouwer's theorem: $\mathcal{E}$ is compact and $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ is continuous, so at least one fix point exists.

3. Interior of $\mathcal{E}$: $\mathcal{E}' = \{\mathbf{u} \in \mathcal{E} : \forall i, u_i > 0\}$.

$$d(\mathbf{u}, \mathbf{v}) = \log \left[\max_i (u_i/v_i) \Big/ \min_i (u_i/v_i) \right]$$

is a distance on $\mathcal{E}'$.

4. d decreases when $\boldsymbol{\psi}$ is applied:

$$d[\boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{u}), \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v})] < d(\mathbf{u}, \mathbf{v})$$

(except if $\mathbf{u} = \mathbf{v}$). $\rightarrow \boldsymbol{\gamma} \circ \boldsymbol{\psi}$ admits at most one fix point in $\mathcal{E}'$.

5. If $k_{ij} > 0$ for all (i, j):

$$\{\mathbf{u} \in \mathcal{E} \setminus \mathcal{E}'\} \quad \Longrightarrow \quad \{\boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{u}) \in \mathcal{E}'\}.$$

$\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ (and therefore for $\boldsymbol{\psi}$) admits one unique fix point toward which the algorithm converges.

Estimation a (and h)

Analogy with EM. a could be estimated iteratively:

$$\widehat{a}^h = \frac{1}{n}\sum_i \widehat{\tau}_i^h$$

but

$$\widehat{\boldsymbol{\tau}} = (\ 1 \quad \ldots \quad 1\ ), \qquad \widehat{a} = 1$$

is a fixed point of this algorithm.

Remark. For a given a, there is a unique $\widehat{\boldsymbol{\tau}}$.

In some sense, a is the unique parameter of the problem.

Linear regression. a may be estimated in an independent way. Ex: linear regression

$$\widehat{a} = \arg\min_a \sum_{i:P_i \geq t} \left\{ \widehat{G}(x) - [(1-a)\Phi(x) + b] \right\}^2 .$$

Cross-validation. a (and h) can also be estimated as follows

1. Split the dataset $\mathcal{D}$ into V subsets $\mathcal{D}_1, \ldots, \mathcal{D}_V$.
   Typically, $V = 5$ or 10.

2. For $v = 1 \ldots V$

   - estimate f and g with the data from $\mathcal{D} \setminus \mathcal{D}_v$ $(\rightarrow \widehat{f}_v, \widehat{g}_v)$,

   - calculate
   $$\mathcal{L}_{CV}(\mathcal{D}; a) = \frac{1}{V} \sum_v \sum_{i \in \mathcal{D}_v} \log \widehat{g}_v(x_i).$$

3. Maximize $\mathcal{L}_{CV}$ (numerically):

$$\widehat{a} = \arg\max_a \mathcal{L}_{CV}(\mathcal{D}; a).$$

FDR and local FDR estimation

Definition. Recall that, when the i tests with smallest p-values are declared positive ($t = P_{(i)}$, $R(t) = i$):

$$FDR_{(i)} = \mathbb{E}[FP(t)/i], \qquad FNR_{(i)} = \mathbb{E}[FN(t)/(1-i+1)]$$

These definitions may be rephrased in terms of mixture model.

$$FDR_{(i)} = \frac{1}{i} \sum_{j:P_j \leq P_{(i)}} (1-\tau_j), \qquad FNR_{(i)} = \frac{1}{n-i+1} \sum_{j:P_j \leq P_{(i)}} \tau_j.$$

The local FDR is $\ell FDR_i = 1 - \tau_i$.

Estimation. We get the natural estimates:

$$\widehat{FDR}_{(i)} = \frac{1}{i} \sum_{j \leq i} (1-\widehat{\tau}_j), \qquad \widehat{FNR}_{(i)} = \frac{1}{n-i+1} \sum_{j>i} \widehat{\tau}_j. \qquad \widehat{\ell FDR} i = 1 - \widehat{\tau}_i.$$

Applications

Probit transform

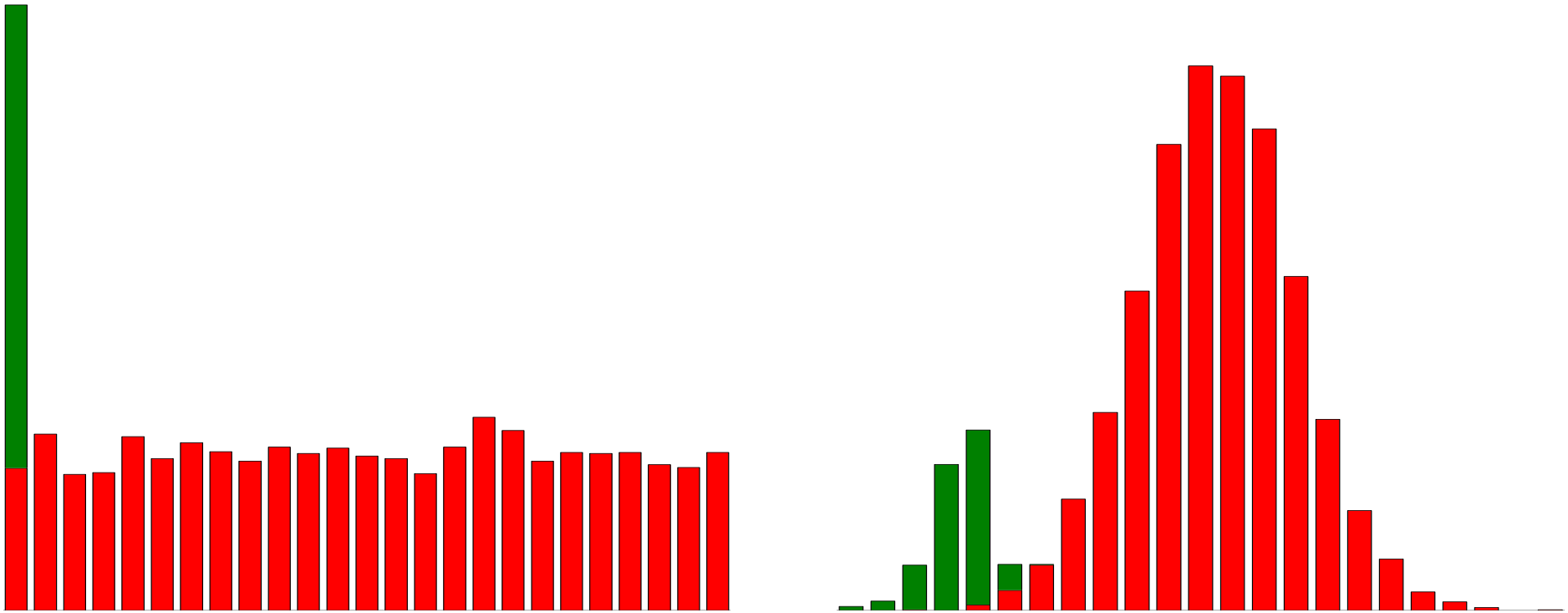
$P_i \in [0,1]$

$X_i = \Phi^{-1}(P_i) \in \mathbb{R}$

(Efron, JASA, 2005)

$\phi = \mathcal{U}_{[0;1]}$

$\phi = \mathcal{N}(0,1)$

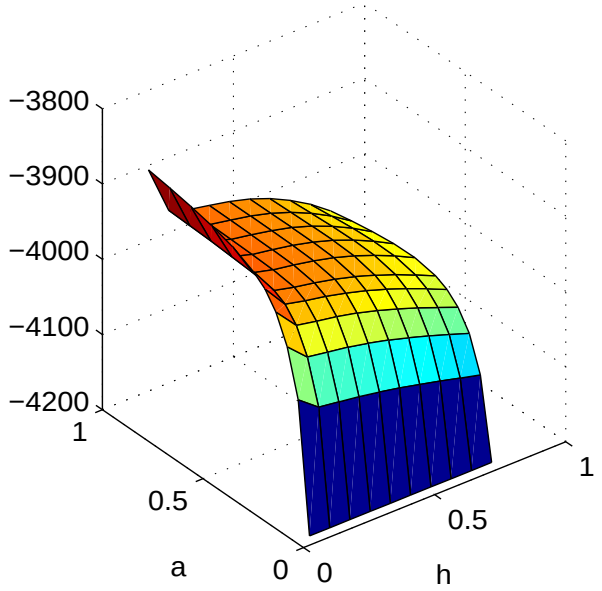


Interest of cross-validation.

Hedenfalk data. Comparison of 2 breast cancers (BRCA1 / BRCA2): $n = 3226$ genes, Epanechnikov kernel, Cochran test

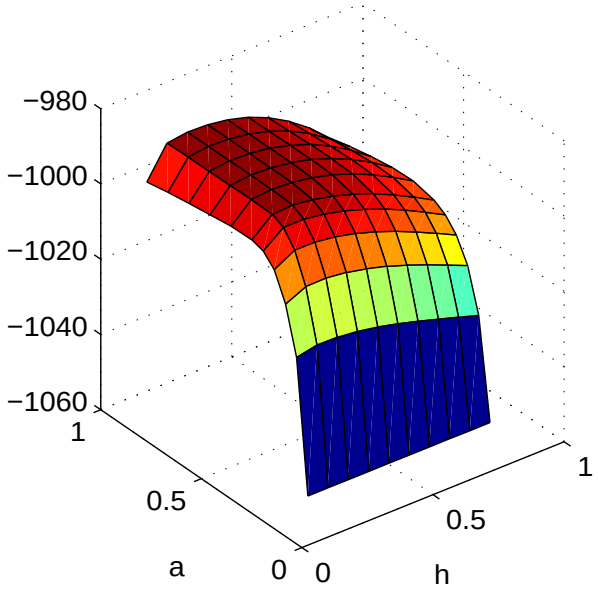
log-likelihood $\mathcal{L}(a, h)$, $(V = 5)$

training set

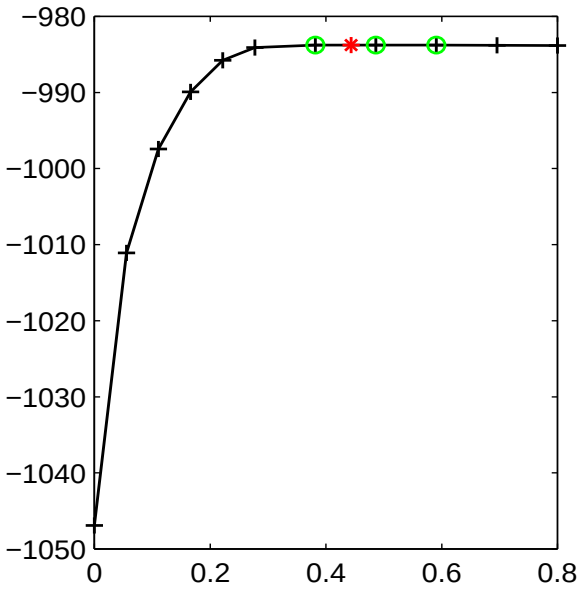


$\widehat{a} \to 1, \quad \widehat{h} \to 0$

test set: $\mathcal{L}_{CV}$



$\widehat{h} = 0.177$

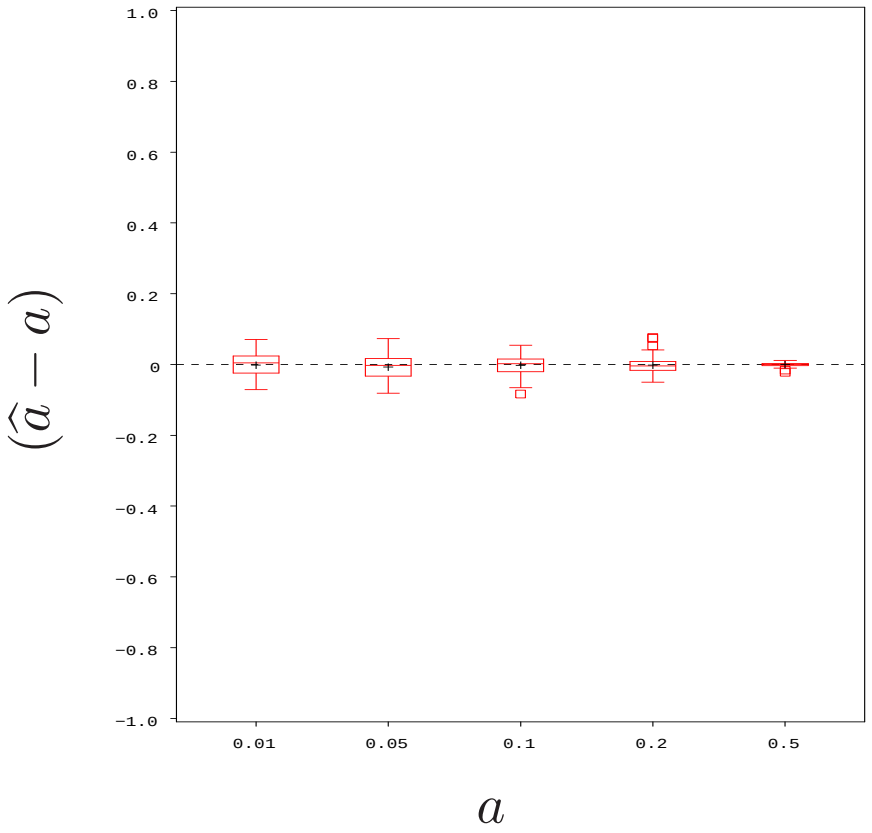


$\widehat{a} = 0.443$

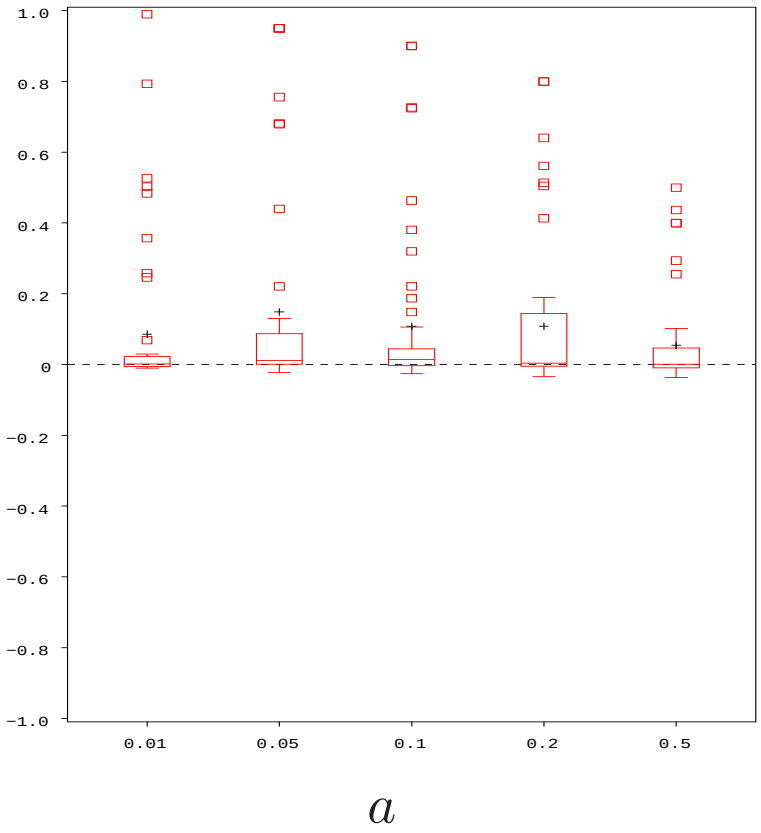
Estimation of a

50 simulations.

Linear regression ($t = 1/2$)

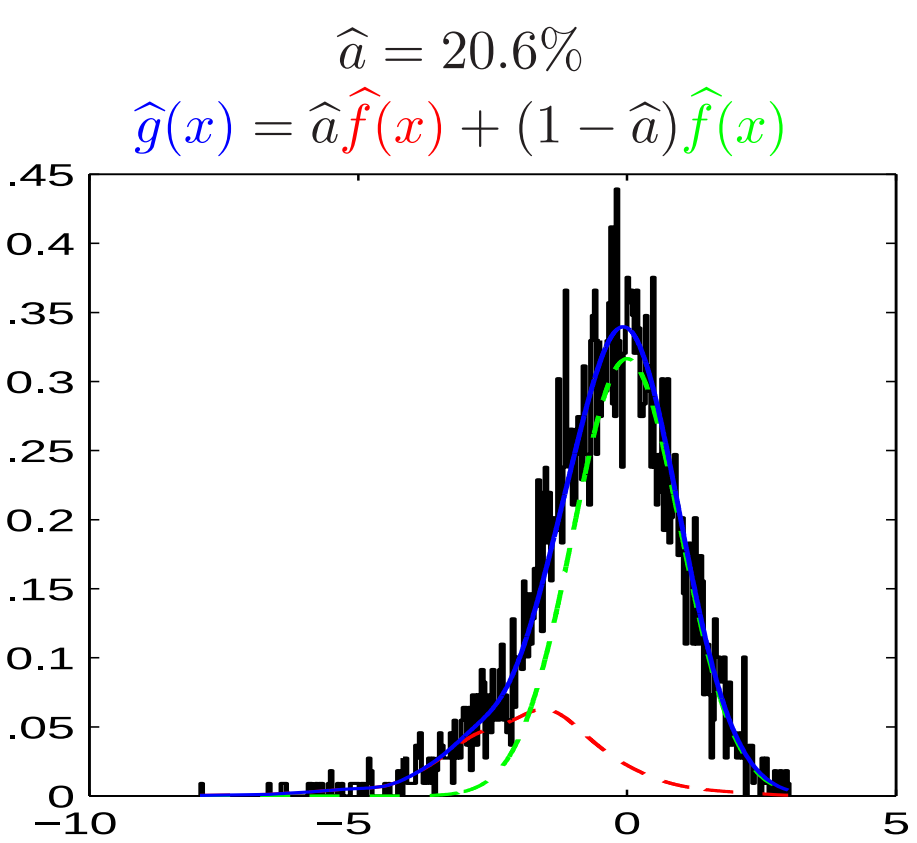


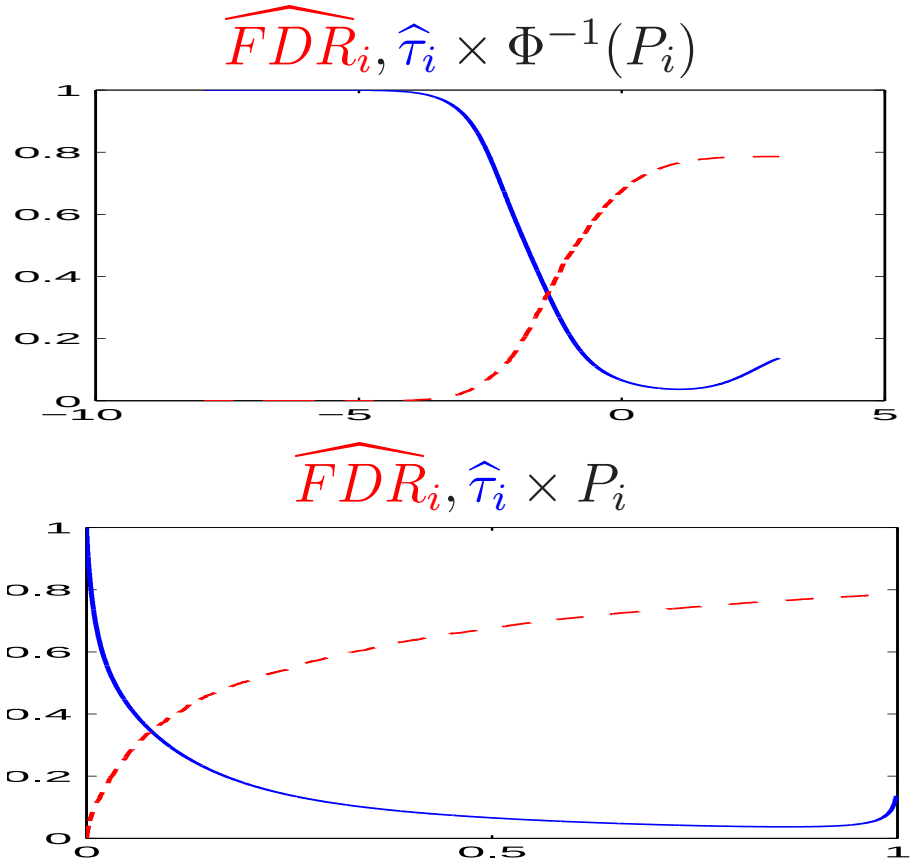
Cross-validation ($\max_a \mathcal{L}_{CV}$)



Hedenfalk data

Student t-test with homogenous variance $\sigma_g = \text{cst}$. Gaussian kernel.



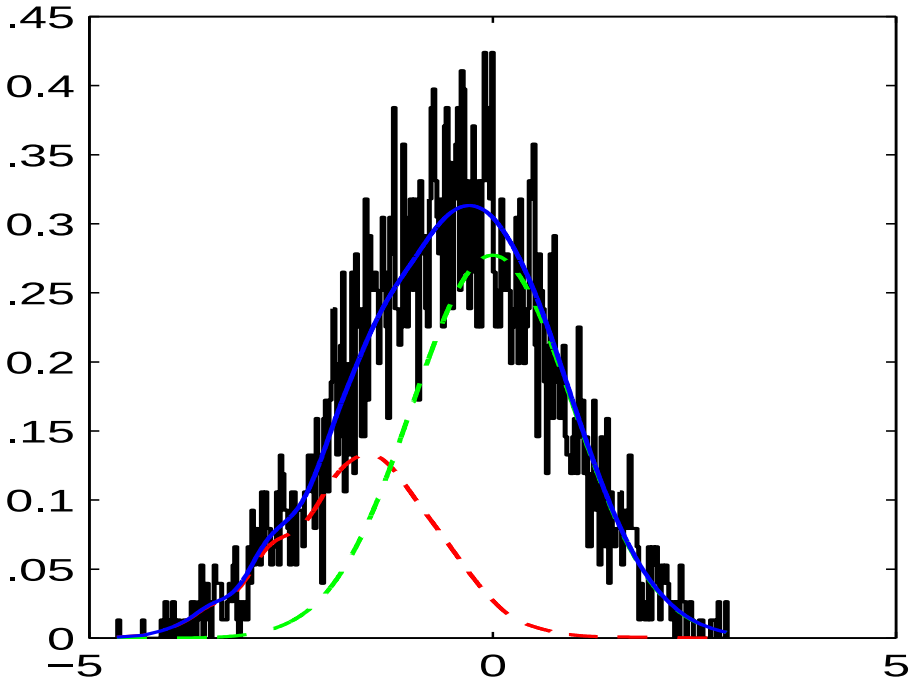


$\mathbf{H}_0$ (negative) p-values are not uniformly distributed over [0, 1]. The non-parametric part is contaminated by this departure of the nu.

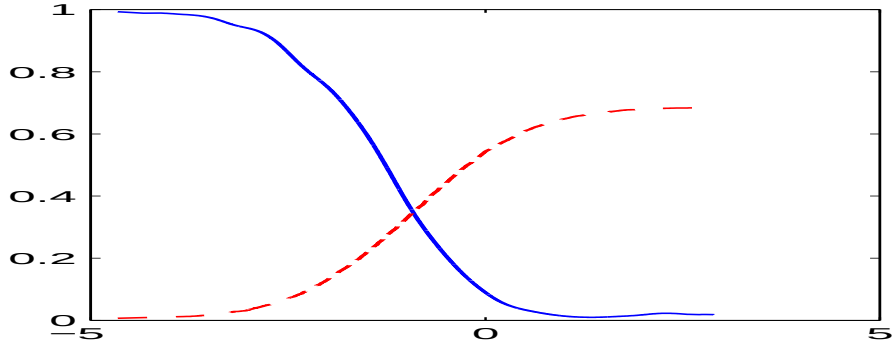
Variance modeling $K = 5$ groups of variances.

$\widehat{a} = 30.5\%$

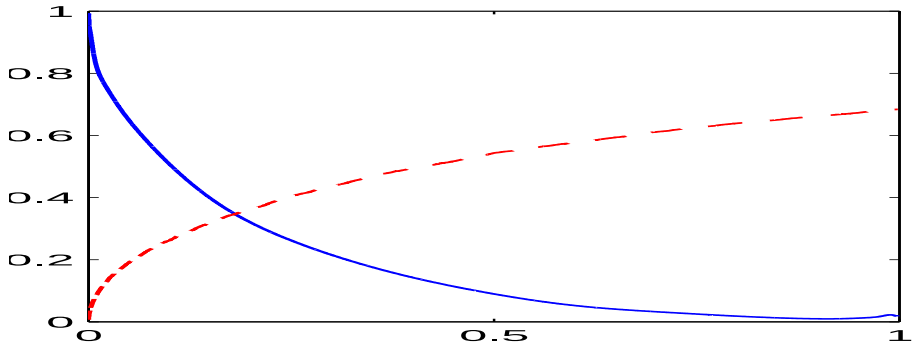
$\widehat{g}(x) = a\widehat{f}(x) + (1-a)\widehat{f}(x)$



$\widehat{FDR}_i, \widehat{\tau}_i \times \Phi^{-1}(P_i)$



$\widehat{FDR}_i, \widehat{\tau}_i \times P_i$



$\widehat{FDR}_{(i)}$	i	$P_{(i)}$	$\widehat{\tau}_{(i)}$	$\widehat{FNR}_{(i)}$
1%	4	$2.5\ 10^{-5}$	0.988	31.5%
5%	142	$3.1\ 10^{-3}$	0.914	28.7%
10%	296	$1.3\ 10^{-2}$	0.798	25.7%

$\widehat{FDR}_{(i)} = \widehat{FNR}_{(i)} = 19.7\%$ for $(i) = 633, P_{(i)} = 5.4\%, \widehat{\tau}_{(i)} = 43.5\%$.