A semi-parametric approach for mixture models, application to local False Discovery Rate estimation

Jean-Jacques Daudin, $AgroParisTech$/INRA,
joint work with
S. Robin, A. Bar-Hen, L. Pierre (Univ. Paris X)

Mixture model

two-populations

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

- probability density function ϕ is known
- probability a is unknown
- probability density function f is unknown.

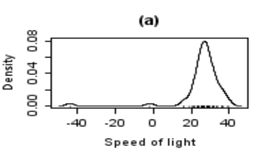
Applications

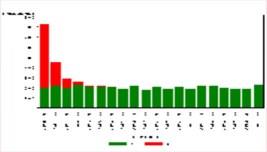
- Contamination problems,
  - distribution ϕ is known,
  - contamination distribution f is unknown,
  - proportion a of contamination is unknown.
- Multiple testing problems
  - p-values under H_0 are uniformly distributed on $[0, 1]$, ϕ is the uniform distribution,
  - distribution of the p-values associated to H_1 is unknown,
  - proportion a of observations under H_1 is unknown.

Data set relating to speed of light

measurements made by Simon Newcomb

(in Gelman et al in Bayesian Data Analysis (2004))





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

Idea Build a kernel nonparametric estimate of f using the information we have on ϕ

Issue

- It is easy to build a kernel density of the overall distribution g, but that is not what we want to do
- we want to build a kernel estimate of f, so we need to know which observations are generated under f.
- this information is not available...

Solution Estimate the probability for each observation of being generated under f (or under ϕ).

Basic relation 1

- Consider an observation x
- Assume that f and a are known,

The probability $\tau(x)$ that this observation has been generated under f is

$$\tau(x) = \frac{af(x)}{g(x)} = \frac{af(x)}{af(x) + (1-a)\phi(x)} \ldots$$

But f and a are unknown, we just wanted to estimate them !

Basic relation 2

The standard kernel estimate of f is

$$\widehat{f}(x) = \left[\sum_i Z_i k_i(x)\right] \Big/ \sum_i Z_i \,.$$

where

- k is a kernel pdf
- $k_i(x) = k[(x - x_i)/h]/h$
- h is the bandwidth of the kernel
- Z_i is one if the data x_i comes from f and 0 otherwise.

Basic relation 2

$$\widehat{f}(x) = \left[\sum_i Z_i k_i(x)\right] \bigg/ \sum_i Z_i \; .$$

can not be directly used since the $\{Z_i\}$ are unknown.
We replace them with their conditional expectation given the data $\{x_i\}$ (i.e. the posterior probabilities) $\mathbb{E}(Z_i \mid x_i) = \tau(x_i)$
We get the following estimate for f:

$$\widehat{f}(x) = \left(\sum_i \tau(x_i) k_i(x)\right) \bigg/ \sum_i \tau(x_i) \; .$$

This estimate is a *weighted kernel estimate* where each observation is weighted according to its posterior probability to be issued from f.

Consistency constraint

Assume a is known. A consistent estimate of f must satisfy the two relations :

- $$\widehat{\tau}(x) = \frac{a\widehat{f}(x)}{a\widehat{f}(x) + (1-a)\phi(x)}.$$
- $$\widehat{f}(x) = \left(\sum_i \widehat{\tau}(x_i) k_i(x)\right) \Big/ \sum_i \widehat{\tau}(x_i) .$$

Two questions

- How many solutions to the consistency constraint : 0, 1 or > 1?
- If the solution is unique, find an algorithm to obtain it

Main result

- Under quite general conditions concerning the kernel function k and the known pdf ϕ,
- for given a, and h and a given sample $(x_i, i = 1, n)$,

$$\boxed{\text{there is a unique solution for } \widehat{f} \text{ (and } \widehat{\tau}(x)\text{).}}$$

This solution is given by a fixed-point algorithm.

Fixed-point equation(1)

$$\widehat{\tau}(x) = \frac{a\widehat{f}(x)}{a\widehat{f}(x) + (1-a)\phi(x)}.$$

$$\widehat{f}(x) = \left(\sum_i \widehat{\tau}(x_i) k_i(x)\right) \Bigg/ \sum_i \widehat{\tau}(x_i)$$

$$\widehat{\tau}(x) = \frac{a\frac{\sum_i \widehat{\tau}(x_i) k_i(x)}{\sum_i \widehat{\tau}(x_i)}}{a\frac{\sum_i \widehat{\tau}(x_i) k_i(x)}{\sum_i \widehat{\tau}(x_i)} + (1-a)\phi(x)}.$$

Fixed-point equation(2)

$(\tau = \tau(x_i),\ \ i = 1:n)$ must satisfy the fixed-point equation

$$\hat{\tau} = \psi(\hat{\tau})$$

where ψ maps $\mathbb{R}^n$ into $\mathbb{R}^n$:

$$\text{For all } \mathbf{u} = (u_1 \dots u_n) \in \mathbb{R}^n : \psi_j(\mathbf{u}) = \frac{\sum_i u_i b_{ij}}{\sum_i u_i b_{ij} + \sum_i u_i},$$

with

$$b_{ij} = \frac{a}{1-a} \frac{k_i(x_j)}{\phi(x_j)}.$$

Theorem

Theorem

If all coefficients b_{ij} are positive, the function ψ has a unique fixed point $\mathbf{u}^*$ and the sequence $\mathbf{u}^{\ell+1} = \psi(\mathbf{u}^\ell)$ converges towards it for any initial value $\mathbf{u}^0$.

Proof.

Rather technical:

- decomposition of ψ as $\psi = \alpha \circ \beta \circ \gamma$
- Brouwer's theorem
- the distance between two points strictly decreases when the function $\gamma \circ \psi$ is applied.
- The condition on b may be relaxed so that non compact kernels are included.

Estimation of a and h

The bandwidth h is obtained by V-fold cross-validation. The following estimate for a is given in the literature in the case of the multiple testing problem:
if the support of the distribution f has an upper bound (typically, $(-\infty, \lambda]$), an unbiased estimate of a can be proposed: for $x > \lambda$, $F(x) = 1$, the mixture cdf becomes

$$G(x) = a + (1 - a)\Phi(x),$$

where G and Φ are the respective cdfs corresponding to g and ϕ.

$$\widehat{a} = \frac{\widehat{G}(\lambda) - \Phi(\lambda)}{1 - \Phi(\lambda)}$$

where $\widehat{G}$ is the empirical cdf of X.

Application to multiple testing

Local FDR

Defined by Efron(2001) in the context of the multiple testing procedure.

It gives the probability for a given observation to be a false positive

In a mixture framework, a natural way to define the local FDR is to consider the posterior probability

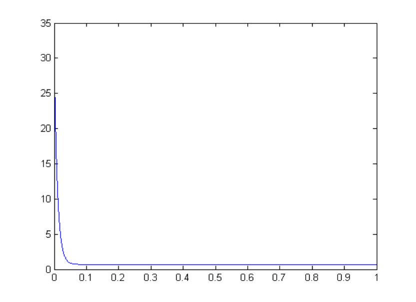
$$\ell\mathrm{FDR}(x) = \Pr\{Z_i = 0 \mid X_i = x\} = 1 - \tau(x).$$

Our kernel nonparametric estimate of f gives directly τ and thus ℓFDR.

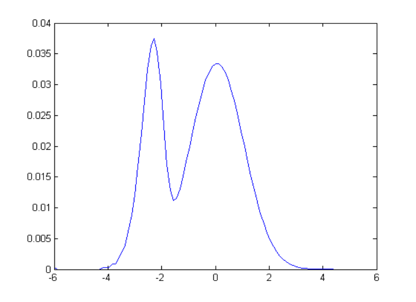
Probit transformation

f : exponential density with mean 0.01 and $a = 0.3$

raw scale

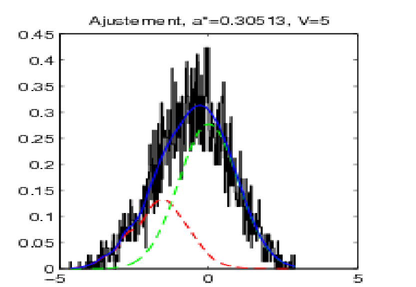


probit scale



Example: Hedenfalk's data, estimation of a, f and τ

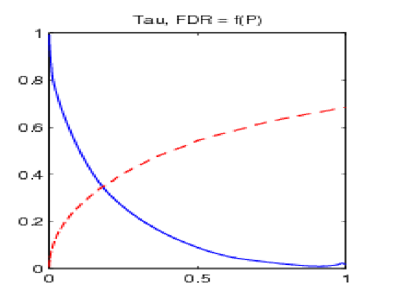
Comparison of 2 breast cancers (BRCA1 / BRCA2), $n = 3226$ genes



black lines : empirical

red : f, green : ϕ, blue : g

x-axis : probit scale



blue curve : $\hat{\tau}(x)$

red : $\widehat{\text{FDR}}(x)$

x-axis : raw scale

x-axis : P−values for H_0 = {gene is not differentially expressed between the 2 conditions }

Example: Hedenfalk's data, control of the FDR

$$\widehat{\mathrm{FDR}}(x_i) = \frac{1}{i}\sum_{j=1}^{i}(1-\hat{\tau}(x_j)), \qquad \widehat{\mathrm{FNR}}(x_i) = \frac{1}{n-i}\sum_{j=i+1}^{n}\hat{\tau}(x_j)$$

$\widehat{\mathrm{FDR}}(x_{(i)})$	i	$\mathrm{P}_{(i)}$	$\hat{\tau}(x_{(i)})$	$\widehat{\mathrm{FNR}}(x_{(i)})$
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%

Table: Number of positive genes for some pre-specified values of the FDR

Methods compared

LocalFDR Efron(2004): mixture model on the probit transformation of the p-values, `locfdr` package of R version 1.3.

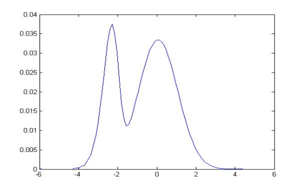
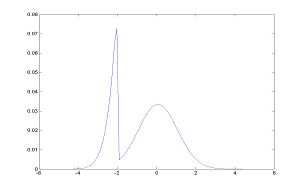
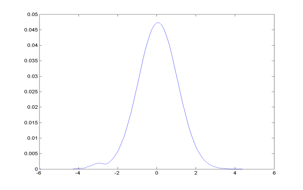
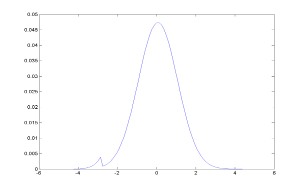
2Gmixt McLachlan(2006): two components gaussian mixture model on the probit transformation of the p-values

SPmixt semi-parametric mixture model on the probit transformation of the p-values

Simulation experiment

Number of simultaneous tests	1000
a	0.01, 0.05, 0.1, 0.3
shape of f	exponential and uniform distributions
mean of f	0.001 and 0.01
Number of simulations	500

Examples of mixtures simulated (probit scale)

f: exponential pdf	f: uniform pdf	μ	a
		0.01	0.3
		0.001	0.01

Criteria for comparison

$$\mathrm{RMSE}^s_m(a,f) = \sqrt{\frac{1}{n}\sum_i \left(\hat{\tau}^s_{m,i} - \tau_i\right)^2}$$

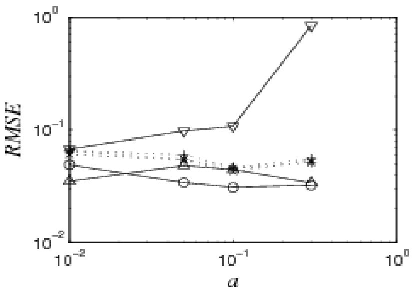
$$\mathrm{RMSE}_m(a,f) = \frac{1}{S}\sum_s \mathrm{RMSE}^s_m(a,f)$$

- s simulation number s ($s = 1..S$)
- τ_i the posterior probability for the ith p-value

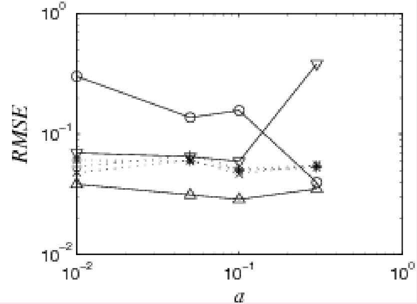
The quality of the estimates provided by method m in the configuration (a,f) is measured by the mean $\mathrm{RMSE}_m(a,f)$.

Simulation results ($f \sim exp(\frac{1}{\mu})$)

$\mu = 0.001$



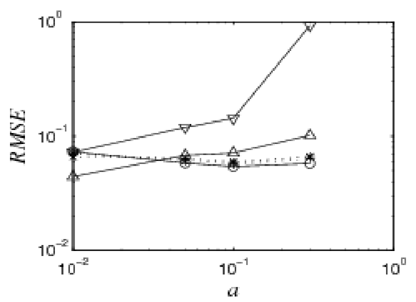
$\mu = 0.01$



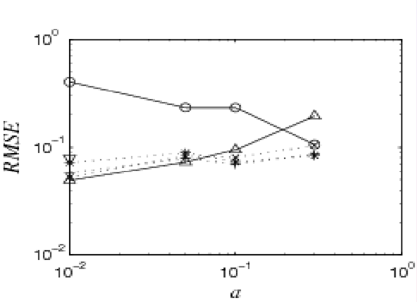
'∇'= default localFDR '$\triangle$'=localFDR-$\mathcal{N}(0,1)$ '$\circ$'= 2Gmixt
'+'= SPmixt with $h = 0.1$ '$\times$'= SPmixt with $h = 0.2$
'$*$'= SPmixt with h fitted using 2-fold cross-validation

Simulation results ($f \sim U[0, 2\mu]$)

$\mu = 0.001$



$\mu = 0.01$



'∇'= default localFDR   '$\triangle$'=localFDR-$\mathcal{N}(0,1)$ '$\circ$'= 2Gmixt

'+'= SPmixt with $h = 0.1$ '$\times$'= SPmixt with $h = 0.2$

'$*$'= SPmixt with h fitted using 2-fold cross-validation

Conclusions

- The weighted kernel compares favorably with competitors
- there is very few information about f, and n is large in multiple testing context $\rightarrow$ nonparametric density estimates are attractive
- weighted nonparametric density estimates : an emerging field
- need more work to obtain simultaneous estimates for a and f in place of the present two stages method.