A semi-parametric approach for mixture models: Application to local FDR estimation

by

Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin

Research Report No. 3
January 2007

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
`http://genome.jouy.inra.fr/ssb/`

A semi-parametric approach for mixture models: Application to local FDR estimation

Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin
INA-PG / INRA Biométrie, 16 rue Claude Bernard, 75005 Paris, France
Laurent Pierre
Université Paris X, 200 avenue de la République, 92001 Nanterre Cedex

Abstract

In this article we propose a procedure to estimate a two-components mixture model where one component is known. The unknown part is estimated with a weighted kernel function. The weights are defined in an adaptative way. We prove the convergence and unicity of our estimation procedure. Using simulations, we compared the proposed procedure with two classical approaches. We also applied our results to multiple testing procedure to estimate the posterior population probabilities and the local FDR.

Key words: FDR, Mixture model, Multiple testing procedure, Semi-parametric density estimation.

1 Introduction

We consider a mixture model with two-populations

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

where the probability density function ϕ is known, the probability a is unknown and the probability density function f is completely unknown.
This model appears in at least two contexts:

- in contamination problems, a distribution ϕ for which reasonable assumptions can be made is contaminated by an arbitrary distribution f ([McLachlan and Peel (2000)]).
- in multiple testing problems (microarrays analysis, neuro-imaging) the p-values under H_0 are uniformly distributed on $[0,1]$ while the distribution of the p-values associated to H_1 is unknown. In this setting, ϕ is the uniform distribution.

In this paper we propose to use a nonparametric estimate of f (using a weighted kernel function and the information we have on ϕ) and apply it in the framework of multiple testing. However, the proposed method is general and may be used in any context which may be modelled by (1).

The idea to mix parametric and nonparametric estimates is not new. [Olkin and Spiegelman (1987)] proposed to use a linear combination of a parametric estimate and a nonparametric estimate, [Hjort and Glad (1995)] proposed to update parametric estimate by nonparametric correction functions. The reverse idea using properties of the exponential family was developed by [Efron and Tibshirani (1996)]. [Priebe and Marchette (2000)] and [Di Marzio and Taylor (2004)] proposed to use parametric estimates for the weights of kernel density estimation.
Using projection pursuit density estimation framework, [Hoti and Holmström (2004)] proposed to estimate a mixture of normal densities with kernels functions. However, the idea of using a nonparametric estimate for f in model (1) is new.

In the framework of multiple testing, mixture models have already been proposed. [Efron *et al.* (2001)] used model (1) to estimate the local FDR defined as the posterior probability of population f, derived from the mixture model

$$\tau(x) = af(x)\,/g(x)\ . \tag{2}$$

Parametric models have been used with Beta distribution for the p-values (see for example [Allison *et al.* (2002)], [Pounds and Morris (2003)], [Liao *et al.* (2004)]) or gaussian distribution of the probit transformation of the p-values ([McLachlan *et al.* (2006)]).

The general approach of our work and the main result are presented in Section 2. Practical issues are discussed in Section 3. Application to the multiple testing procedure and estimation of the local FDR is studied in Section 4. We compared our method to those proposed by [Efron (2004)] and [McLachlan *et al.* (2006)] on simulated data in section 5. The last section is devoted to the application of the proposed procedure to the classical Hedenfalk dataset ([Hedenfalk *et al.* (2001)]).

2 Estimation of the unknown density

2.1 Kernel estimate

Since f is completely unspecified, it has to be estimated in a non-parametric way. Let k denote an arbitrary cdf (called kernel function), 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 $k_i(x) = k[(x - x_i)/h]/h$, h is the bandwidth of the kernel and Z_i is one if the data x_i comes from f and 0 otherwise.

This estimate can not be directly used since the $\{Z_i\}$ are unknown. We propose to replace them with their conditional expectation given the data $\{x_i\}$ that are equal to the posterior probabilities: $\mathbb{E}(Z_i \mid x_i) = \tau(x_i)$ as defined in equation 2. We get the following estimate for f:

$$\widehat{f}(x) = \left(\sum_i \tau(x_i) k_i(x)\right) \Big/ \sum_i \tau(x_i) \ . \tag{3}$$

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

2.2 Estimation of the posterior probabilities

The conjunction of (3) and (2) implies that the vector $\widehat{\boldsymbol{\tau}}$ containing the estimated posterior probabilities $\tau(x_i)$ must satisfy

$$\widehat{\boldsymbol{\tau}} = \boldsymbol{\psi}(\widehat{\boldsymbol{\tau}}) \tag{4}$$

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

$$\text{For all } \mathbf{u} = (u_1 \dots u_n) \in \mathbb{R}^n : \psi_j(\mathbf{u}) = \frac{\sum_i u_i b_{ij}}{\sum_i u_i b_{ij} + \sum_i u_i}, \qquad \text{with } b_{ij} = \frac{a}{1-a}\frac{k_i(x_j)}{\phi(x_j)}. \tag{5}$$

$\widehat{\boldsymbol{\tau}}$ must therefore be a fixed point of the function $\boldsymbol{\psi}$.

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

The proof of this theorem is based on the decomposition of $\boldsymbol{\psi}$ as $\boldsymbol{\psi} = \boldsymbol{\alpha} \circ \boldsymbol{\beta} \circ \boldsymbol{\gamma}$ where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are functions mapping from $\mathbb{R}^n$ into $\mathbb{R}^n$:

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

$\boldsymbol{\gamma}$ is actually the projection onto the simplex $\mathcal{E} = \{\mathbf{u} \in \mathbb{R}^n : \sum_i u_i = 1\}$.
The proof requires the three following lemmas, the proofs are given in Appendix.

Lemma 1 *$\mathbf{u}^*$ is a fixed point of $\boldsymbol{\psi}$ if and only if $\mathbf{v}^* = \boldsymbol{\gamma}(\mathbf{u}^*)$ is a fixed point of $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$.*

Lemma 2 *Consider the interior $\mathcal{F}$ of the simplex $\mathcal{E}$: $\mathcal{F} = \{\mathbf{u} \in \mathcal{E} : \text{ For all } i, u_i > 0\}$. The function d mapping $\mathcal{F} \times \mathcal{F}$ into $\mathbb{R}+$:*

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

is a distance.

Lemma 3 *For any $\mathbf{v}$ and $\mathbf{w}$ in $\mathcal{F}$, we have*

$$d(\boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v}), \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{w})) < d(\mathbf{v}, \mathbf{w})$$

if $\mathbf{v} \neq \mathbf{w}$, and $d(\boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v}), \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{w})) = d(\mathbf{v}, \mathbf{w}) = 0$ otherwise.

Proof of Theorem 1. Thanks to Lemma **1**, we can restrict the proof to the study of the convergence of the sequence $\mathbf{v}^{\ell+1} = \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v}^{\ell})$ in the simplex $\mathcal{E}$. Since $\mathcal{E}$ is a compact and $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ is continuous, Brouwer's theorem insures that $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ admits at least one fixed point in $\mathcal{E}$.

Furthermore, since every b_{ij} is strictly greater than zero, the image $\boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v})$ of any element $\mathbf{v}$ of $\mathcal{E}$ can not have any null coordinate. That is: the function $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ sends the elements of the border of $\mathcal{E}$ into its interior $\mathcal{F}$. So the fixed points of $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ necessarily belong to $\mathcal{F}$.

Lemma **3** proves that $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ admits at most one fixed point since the 2 fixed points case would contradict the lemma for $\mathbf{v} \neq \mathbf{w}$. This implies that there exist a unique fixed point. Finally, Lemma **3** says that the distance d (Lemma **2**) strictly decreases when the function $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ is applied. This shows that the iteration of the function $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ necessarily converges to its unique fixed point and achieves the proof. ■

Hypothesis on the b_{ij}. This hypothesis may be relaxed. The second argument of the proof still holds if the function $\boldsymbol{\psi}$ sends any element of the border of $\mathcal{E}$ into the interior in a finite number of iterations. In terms of kernel estimate, the convergence is therefore guaranteed for kernel with non-compact support (such as the Gaussian kernel), or if no observation x_i is isolated within its kernel, i.e. if

$$\text{For all } i, \text{ it exists } j \neq i : k_i(x_j) > 0.$$

Value of a and h. These results are conditional on a and h, which is a necessary condition for theorem 1.

2.3 Estimation algorithm

The iteration of the function $\boldsymbol{\psi}$ can be decomposed in 3 sub-steps, as exposed in Algorithm 1.

Algorithm 1

```
Initialization:                          Set $\widehat{\tau}^0(x_i)$ to 1 for the proportion $a$ of the smallest
                                         $x_i$ and to 0 for the remaining.

Step $\ell$   estimation of $f$:              $\widehat{f}^{(\ell)}(x) = \sum_i \widehat{\tau}^{(\ell-1)}(x_i)k_i(x) \,/ \sum_i \widehat{\tau}^{(\ell-1)}(x_i)$

         estimation of $g$:              $\widehat{g}^{(\ell)}(x) = a\widehat{f}^{(\ell)}(x) + (1-a)\phi(x)$

         update of $\{\tau(x_i)\}$:          $\widehat{\tau}^{(\ell)}(x_i) = a\widehat{f}^{(\ell)}(x_i) \,/ \widehat{g}^{(\ell)}(x_i)$

Stopping rule                            Stop when $\max_i |\widehat{\tau}^{(\ell)}(x_i) - \widehat{\tau}^{(\ell-1)}(x_i)|/\widehat{\tau}^{(\ell-1)}(x_i) < \varepsilon$.
```

where $\widehat{f}^{(\ell)}$, $\widehat{g}^{(\ell)}$ and $\widehat{\tau}^{(\ell)}(x_i)$ denote the estimates of f, g and $\tau(x_i)$ after step ℓ.

Connexion with the E-M algorithm. Algorithm 1 has some Expectation-Maximization (E-M) flavor. Actually the updating of the $\widehat{\tau}(x_i)$ is equivalent to the E step. Moreover, considering the $\{k_i(x)\}$ as data, $\widehat{f}(x)$ can be seen as an average of them, so the updating of $\widehat{f}$ may look like an M step.
However, this comparison is not valid since kernel estimates do not aim at maximizing the likelihood of the data (like E-M does), but typically to minimize the norm of $(\widehat{f} - f)$. Therefore, Algorithm 1 can not be justified in the standard E-M framework for mixture models. The algorithm does not optimize any given criterion. The relation in equation 4 insures self-consistency of $\widehat{\tau}$

3 Estimation of the proportion and bandwidth

3.1 Simultaneous estimation of a and f

The analogy with the E-M algorithm suggests to estimate a using a modified version of Algorithm 1, where $\widehat{a}$ is updated at each step:

$$\widehat{a}^{(\ell)} = \frac{1}{n} \sum_i \widehat{\tau}^{(\ell)}(x_i).$$

However, it can be easily seen that the solution $\widehat{a} = 1$ and $\widehat{\tau}(x_i) = 1$, for all i, is a fixed point of this modified algorithm. This solution corresponds to the standard kernel estimate of g (not of f). This property can be interpreted as an over-fitting trend.

3.2 Estimation of a

The estimation of a is a difficult task that can not be achieved by the algorithm proposed in the preceding section. In the case where the support of the distribution f has an upper bound (typically, $(-\infty, \lambda]$), two unbiased estimates of a can be proposed. Both come from the observation that, for $x > \lambda$, if $F(x) = 1$, the mixture cdf becomes

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

where G and Φ are the respective cdfs of g and ϕ. In the framework of FDR control, [Storey *et al.* (2004)] proposes

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

where $\widehat{G}$ is the empirical cdf of X. If λ is underestimated, at worst $\widehat{a}$ is underestimated. The authors discuss the performances of these estimates and its sensitivity to the choice of λ. Following the same principle, a can be estimated using a linear least square fit of $\widehat{G}(X_i)$ to $\Phi(X_i)$, that is

$$\widehat{a} = \arg\min_a \sum_{i:X_i > \lambda} (\widehat{G}(X_i) - b - (1-a)\Phi(X_i))^2.$$

where b is a constant. Provided λ exists and is known, both estimates are unbiased. However, both rely on the existence of some additional information about the relative positions of distributions f and ϕ.

3.3 Estimation of h

To estimate the bandwidth h, we propose to use the standard approach ([Silverman (1986)]) based on V-fold cross-validation. We split randomly the data set $\{x_i\}_{i=1..n}$ into V non-overlapping subsets $\mathcal{Y}_1, \dots, \mathcal{Y}_V$, each of size n/V: $\cup_v \mathcal{Y}_v = \{x_i\}_{i=1..n}$. For each $v = 1 \dots V$, we define $\mathcal{X}_v = \cup_{u \neq v} \mathcal{Y}_u$ as the training set, and $\mathcal{Y}_v$ as the test set. We denote $\mathcal{L}(\mathcal{Y}_v; h)$ the log-likelihood of the subset $\mathcal{Y}_v$:

$$\mathcal{L}(\mathcal{Y}_v; h) = \sum_{x_j \in \mathcal{Y}_v} \ln \widehat{g}_v(x_j; h)$$

where $\widehat{g}_v$ is estimated with Algorithm 1 on the training set $\mathcal{X}_v$ with the given window width h. We define the V-fold cross validation log-likelihood as

$$\mathcal{L}_{CV}(h) = \frac{1}{V} \sum_v \mathcal{L}(\mathcal{Y}_v; h).$$

$n^{-1} \left(\sum_{i=1}^n \ln g(x_i) - V\mathcal{L}_{CV}(h) \right)$ is an estimate of the Kullback-Leibler divergence between $\widehat{g}$ and g. This estimated divergence between $\widehat{g}$ and g is minimized when the cross-validation likelihood is maximized, that is for

$$\widehat{h} = \arg\max_h \mathcal{L}_{CV}(h).$$

This optimization can be performed numerically.

4 False positive and negative rates

4.1 Presentation and definitions

Multiple testing is a classical problem for many high-dimensional data sets, since uncorrected testing procedure may lead to many false positives. The breakthrough of technology for image analysis or genomic data has given a new interest for this question. A central problem in multiple testing problems is the control of type I (i.e. false positive) and type II (i.e. false negative) errors. For a given threshold t, we denote

$$\begin{array}{rcll} P(t) & = & \#\{j : X_j < t\} & \text{the number of positives;} \\ FP(t) & = & \#\{j : (X_j < t) \cap (Z_j = 0)\} & \text{the number of false positives;} \\ N(t) & = & \#\{j : X_j \geq t\} & \text{the number of negatives;} \\ FN(t) & = & \#\{j : (X_j \leq t) \cap (Z_j = 1)\} & \text{the number of false positives.} \end{array}$$

The most popular criterion regarding type I errors is the FDR ([Benjamini and Hochberg (1995)]):

$$FDR(t) = \mathbb{E}\left[FP(t)/\max\{P(t), 1\}\right].$$

FDR is the expected proportion of rejections that are incorrect. It is worth noting that a dual quantity of the FDR is the FNR (false non-discovery rate) as defined by [Genovese and Wasserman (2002)]:

$$FNR(t) = \mathbb{E}\left(FN(t)/\max(N(t), 1)\right).$$

In the mixture model framework FDR and FNR play symmetric roles.

More recently, it has been pointed out that, in many multiple testing framework, we need information at the individual level about the probability for a given observation to be a false positive ([Aubert *et al.* (2004)]). This motivated the work of [Storey and Tibshirani (2003)] regarding the q-value. One may notice that the q-value is actually not specific to a given observation since it is computed on all the p-values below a given threshold. In a mixture framework, a natural way to define a 'local FDR' (ℓFDR: [Efron *et al.* (2001)]) is to consider the posterior probability

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

4.2 Estimation

Local FDR. According to the definition given above, a natural estimator of the local FDR for observation i is

$$\widehat{\ell FDR}(x_i) = 1 - \widehat{\tau}(x_i)$$

False positive and negative rates. Following the same approach, we also get:

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

which are unbiased if the posterior probability estimates are unbiased. Since the estimate of $\tau(x_j)$ is proportional to the estimate of a, underestimating a leads to overestimate FDR and underestimate FNR.

We remark that the estimates $\widehat{\ell FDR}$ and $\widehat{FDR}$ are consistent with the definition of ℓFDR in terms of derivative of FDR proposed by [Aubert *et al.* (2004)].

5 Simulation study

We compared our method to those proposed by [Efron (2004)] and [McLachlan *et al.* (2006)] on simulated data. In the following, these methods will be refeered to as 'LocalFDR' and '2Gmixt' (for 2 Gaussian component mixture), respectively. The methods we propose will be denoted by 'SPmixt' (for semi-parametric mixture).

5.1 Simulation design.

We simulated sets of p-values according to the mixture model $g(\cdot) = af(\cdot) + (1-a)\phi(\cdot)$, where ϕ is the uniform distribution over $[0;1]$. This framework is common to the three approaches to be compared. We considered 4 different proportions ($a = 0.01, 0.05, 0.1, 0.3$), 2 different means for the p-values coming from the alternative distribution f ($\mu = 0.01$ and 0.001) and 2 shapes for f (exponential and uniform over $[0; 2\mu]$). For each of the $4 \times 2 \times 2 = 16$ configurations, $S = 500$ samples of size $n = 1000$ were generated.

For each proportion a and distribution f, the posterior probability τ can be computed theoretically for any p-value. To evaluate the performance of each method m in simulation s, we calculated the estimates $\widehat{\tau}_m$ and computed the square root of the mean squared difference between the estimates and the true values

$$RMSE^s_m(a,f) = \sqrt{\frac{1}{n}\sum_i \left(\widehat{\tau}^s_{m,i} - \tau_i\right)^2}, \qquad RMSE_m(a,f) = \frac{1}{S}\sum_s RMSE^s_m(a,f)$$

denoting s the simulation number s ($s = 1..S$) and 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 $RMSE_m(a,f)$.

5.2 Pratical implementation.

For the localFDR method, we used the `locfdr` package of R version 1.3. The complete default options results in many warnings and failures, so we had to fix the `sig0` parameter to 1. In the following, this setting will be refereed to as 'default localFDR'. We also used this method with the `nulltype=0` option, which sets the null distribution to an $\mathcal{N}(0,1)$. In the following, this method will be denoted 'localFDR-$\mathcal{N}(0,1)$'.

For our method, we either fixed the window width h to a given value (0.1 or 0.2) or fitted it using 2-fold cross-validation. We used 2-fold in place of 5-fold (as suggested above) for computation time reasons.

5.3 Results.

Figure 1 displays the $RMSE$ obtained with the different methods under various simulation conditions. We first observe that the result of SPmixt are not very sensitive to the way h is chosen. $RMSE$s are always very similar, whatever the value of h (0.1, 0.2 or 2-fold).

The second comment is that SPmixt provides the most stable and reliable estimates among the considered methods. The default localFDR method provides bad estimates in many situations. It even failed for three values of a in the bottom right plot of Figure 1. No of the three other methods (localFDR-$\mathcal{N}(0,1)$, 2Gmixt, SPmixt) is uniformly the best. 2Gmixt outperforms SPmixt when μ is small (0.001), which corresponds to an easy case where the alternative distribution f is very far from the null one ϕ. 2Gmixt does not perform well when μ is large. LocalFDR-$\mathcal{N}(0,1)$ provides good results when f has an exponential shape, especially when $\mu = 0.01$.

Figure 2 displays the standard deviation of $RMSE^s$ across the simulations. We see that the 2-fold strategy for h induces some variability, that may be reduced using 5-fold cross-validation. We also see that, in terms of stability, SPmixt generally outperforms 2Gmixt. The results provided by localFDR-$\mathcal{N}(0,1)$ strongly depend on the proportion a.

6 Application: Hedenfalk data

[Hedenfalk *et al.* (2001)] compare the gene expression levels measured on patients with two different breast cancer. The dataset consists of 7 BRCA1 patients and 8 BRCA2 patients corresponding to two different gene mutation predisposing to the disease. The total number of genes is $n = 3226$.

Differential analysis. We used a t-test to detect differentially expressed genes. Because of the small number of replicates, the estimate of within group variability appears to very quite poor. This bad estimation is known to have strong consequences on the conclusion. To avoid this problem, we used test statistics and p values P_i computed under the two following hypotheses regarding the variance.

(i) Homogenous variance. The variance of all the genes are all equal to a same variance σ^2.

(ii) Mixture model. Genes are spread into K groups of variance, the variance and proportion of which can estimated using a mixture model ([Delmar *et al.* (2005)]).

Other variance modeling have been proposed: see [Efron *et al.* (2001), Smyth (2004), Rudemo *et al.* (2002)].

Semi-parametric modeling. In this situation, the p-values are expected to have a mixture distribution

$$p_i \sim aF + (1-a)\mathcal{U}_{[0;1]}$$

This mixture is hard to identify because of many p-values very close to 0. Therefore, we used the probit transform suggested by [Efron (2004)], and considered the mixture model on the transformed p-values:

$$X_i = \Phi^{-1}(p_i) \sim aF + (1-a)\Phi$$

were Φ is the cdf of the standard Gaussian distribution.

Results. Figure 3 presents the fit of the semi-parametric mixture model to the histogram of the transformed p-values. We see that in both cases, the distribution f and ϕ strongly overlap. In both case, we used the least-square estimate of a presented in Section 3.2. It resulted in $\widehat{a} = 20.6\%$ in the homogenous variances case, and in $\widehat{a} = 30.5\%$ in the mixture variance case. We set $\lambda = \frac{1}{2}$ and verify, as noted by [Storey *et al.* (2004)], that $\widehat{a}$ is not very sensitive to this choice.

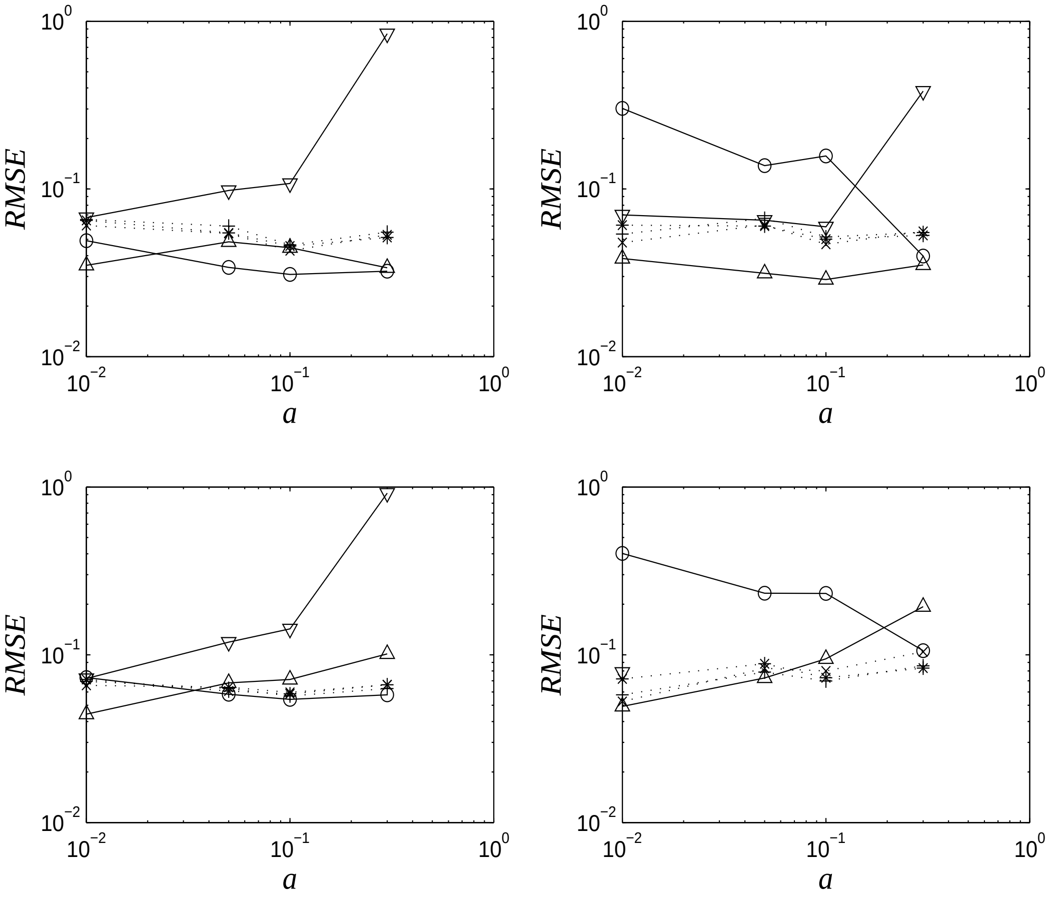

Figure 1: Root Mean Square Error ($RMSE$) between the true posterior probabilities τ and the estimates as a function of the proportion a (log-log scale). Methods: '∇'= default localFDR, '$\triangle$'= localFDR-$\mathcal{N}(0,1)$, '$\circ$'= 2Gmixt, '$+$'= SPmixt with $h = 0.1$, '$\times$'= SPmixt with $h = 0.2$, '$*$'= SPmixt with h fitted using 2-fold cross-validation. Top: exponential shape for f. Bottom: uniform shape. Left: $\mu = 0.001$. Right: $\mu = 0.01$.

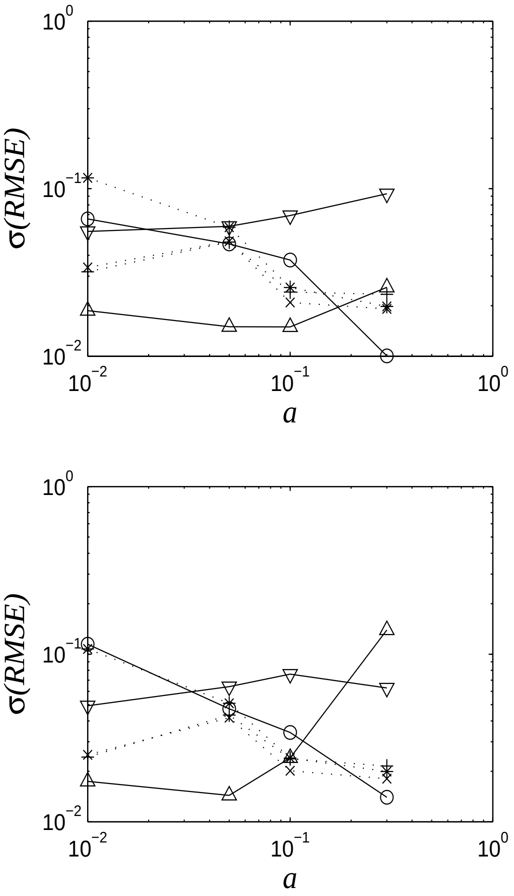

Figure 2: Standard deviation of the $RMSE$ (log-log scale). Methods: '∇'= default localFDR, '$\triangle$'= localFDR-$\mathcal{N}(0,1)$, '$\circ$'= 2Gmixt, '$+$'= SPmixt with $h = 0.1$, '$\times$'= SPmixt with $h = 0.2$, '$*$'= SPmixt with h fitted using 2-fold cross-validation. Top: exponential shape for f. Bottom: uniform shape. Left: $\mu = 0.001$. Right: $\mu = 0.01$.

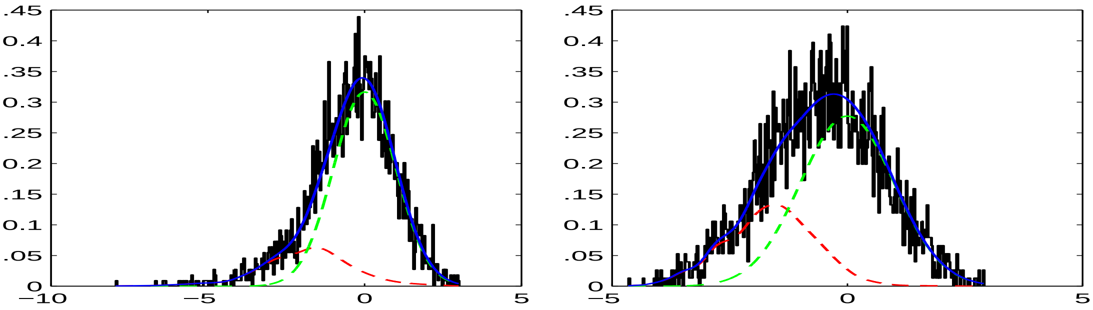

Figure 3: Fit of the semi-parametric mixture model to the transformed p-values for homogenous (left) and mixture (right) gene variances. –: histogram, –: mixture density, - -: $\widehat{a}\widehat{f}$, - -: $(1 - \widehat{a})\phi$.

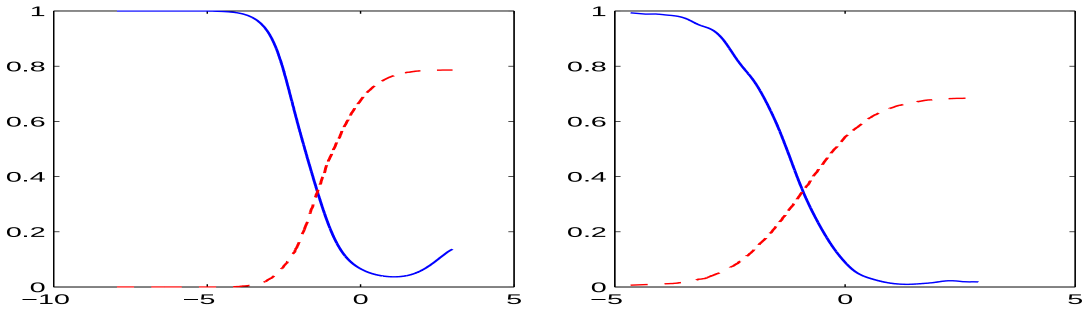

Figure 4: Posterior probabilities $\widehat{\tau}$ ($-$) and false discovery rate $\widehat{FDR}(x_i)$ (- -) as a function of the transformed p-values X_i. Left: homogenous gene variances, right: mixture gene variances.

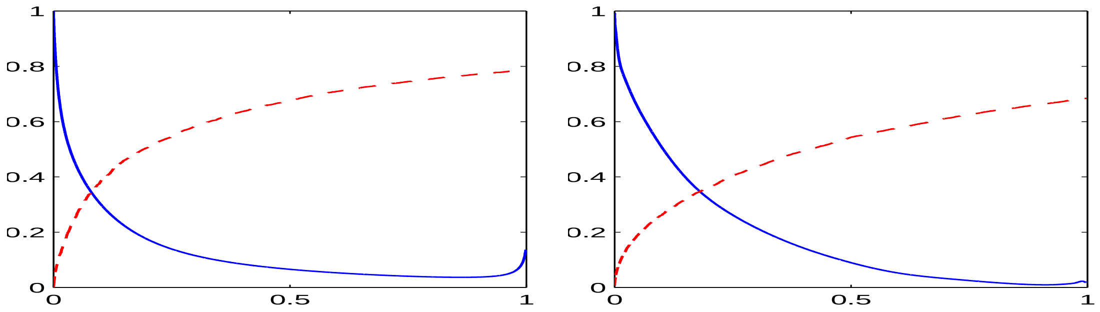

Figure 5: Posterior probabilities $\widehat{\tau}$ ($-$) and false discovery rate $\widehat{FDR}(x_i)$ (- -) as a function of the p-values P_i. Left: homogenous gene variances, right: mixture gene variances.

Figures 4 and 5 present the estimated posterior probabilities $\widehat{\tau}(x_i)$ and $\widehat{FDR}(x_i)$ as a function of X_i and P_i respectively. In the homogeneous variance case, we see that the posterior probabilities first decreases (as expected), and then re-increases on the right part of the plot, which is unexpected. The explanation is that the non-parametric part of the mixture model actually capture a lack of fit of the true distribution of the test statistic to the theoretical distribution under the null hypothesis. This phenomena is strongly reduced by the mixture model for the variances.

Table 1 gives the number of positive genes for some pre-specified values of the FDR. We see that, for small FDR, the minimal posterior probability is still high, which means that all the positive genes can be trusted. We also see that FNR slowly decreases. The estimated FDR and FNR are equal (19.7%) for $i = 633$ positive genes: the corresponding p-value is $P_{[i]} = 5.4\%$, the posterior probability is $\widehat{\tau}(x_{(i)}) = 43.5\%$. This means that, at this point, some of the positive genes are really questionable.

The results in Table 1 are in fair agreement with the results reported in Table 1 of [McLachlan *et al.* (2006)].

$\widehat{FDR}(x_{(i)})$	i	$P_{(i)}$	$\widehat{\tau}(x_{(i)})$	$\widehat{FNR}(x_{(i)})$
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 1: Number of positive genes for some pre-specified values of the FDR

References

[Allison *et al.* (2002)] Allison, D. B., Gadbury, G., Heo, M., Fernandez, J., Lee, C.-K., Prolla, T. A. and Weindruch, R. A. (2002). Mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. and Data Analysis.* **39** 1–20.

[Aubert *et al.* (2004)] Aubert, J., Bar-Hen, A., Daudin, J.-J. and Robin, S. (2004). Determination of the regulated genes in microarray experiments using local FDR. *BMC Bioinformatics.* **5 (125)** 1. http://www.biomedcentral.com/1471-2105/5/125/.

[Benjamini and Hochberg (1995)] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerfull approach to multiple testing. *JRSSB.* **57 (1)** 289–300.

[Delmar *et al.* (2005)] Delmar, P., Robin, S., Le Roux, D. and Daudin, J.-J. (2005). Mixture model on the variance for the differential analysis of gene expression. *J. R. Statist. Soc. C.* **54 (1)** 31–50.

[Di Marzio and Taylor (2004)] Di Marzio, M. and Taylor, C. (2004). Boosting kernel density estimates: A bias reduction technique? *Biometrika.* **91** 226–233.

[Efron (2004)] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99 (465)** 96–104.

[Efron *et al.* (2001)] Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160.

[Efron and Tibshirani (1996)] Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24 (6)** 2431–2461.

[Genovese and Wasserman (2002)] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the fdr procedure. *J. R. Statist. Soc. B.* **64 (1)** 499–518.

[Hedenfalk *et al.* (2001)] Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A. and Trent, J. (2001). Gene expression profiles in hereditary breast cancer. *New Engl. Jour. Medicine.* **(344)** 539–548.

[Hjort and Glad (1995)] Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23 (3)** 882–904.

[Hoti and Holmström (2004)] Hoti, F. and Holmström, L. (2004). A semiparametric density estimation approach to pattern classification. *Comput. Statist. and Data Analysis.* **37** 409–419.

[Liao *et al.* (2004)] Liao, J. G., Lin, Y., Selvanayagam, Z. E. and Weichung, J. S. (2004). A mixture model for estimating the local false discovery rate in dna microarray analysis. *Bioinformatics.* **20 (16)** 2694–2701.

[McLachlan *et al.* (2006)] McLachlan, G., Bean, R.W. and Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics.* **22** 1608-1615.

[McLachlan and Peel (2000)] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley.

[Olkin and Spiegelman (1987)] Olkin, I. and Spiegelman, C. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **82** 858–865.

[Pounds and Morris (2003)] Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics.* **19** 1236–42.

[Priebe and Marchette (2000)] Priebe, C. E. and Marchette, D. J. (2000). Alternating kernel and mixture density estimates. *Comput. Statist. and Data Analysis.* **35 (1)** 43–65.

[Rudemo *et al.* (2002)] Rudemo, M., Lobovkina, T., Mostad, P., Scheidl, S. J., Nilsson, S. and Lindahl, P. (2002), Variance models for microarray data. Technical Report 6, Mathematical Statistics, Chalmers University of Thechnology. `http://www.math.chalmers.se/~rudemo/`.

[Silverman (1986)] Silverman, B. W. (1986). *Density Estimation.* London: Chapman and Hall.

[Smyth (2004)] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarrays experiments *Stat. Appl. Genet. Mol. Biol..* **3 (1)** Article 3

[Storey *et al.* (2004)] Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Statist. Soc. B.* **66 (1)** 187–205.

[Storey and Tibshirani (2003)] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genowide studies. *Proc. Natl. Acad. Sci. USA.* **100 (16)** 9440–45.

Appendix

Proof of Lemma 1. We have $\boldsymbol{\psi} = \boldsymbol{\alpha} \circ \boldsymbol{\beta} \circ \boldsymbol{\gamma}$. Since $\boldsymbol{\gamma}$ is a projection, we have $\boldsymbol{\psi} \circ \boldsymbol{\gamma} = \boldsymbol{\psi}$. So, for $\mathbf{v}^* = \boldsymbol{\gamma}(\mathbf{u}^*)$, we have $\boldsymbol{\psi}(\mathbf{u}^*) = \boldsymbol{\psi}(\mathbf{v}^*)$ and

$$\mathbf{u}^* = \boldsymbol{\psi}(\mathbf{u}^*) \quad \Rightarrow \quad \mathbf{v}^* = \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{u}^*) = \boldsymbol{\gamma} \circ \boldsymbol{\psi} \circ \boldsymbol{\gamma}(\mathbf{u}^*) = \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v}^*).$$

Conversely, let $\mathbf{v}^*$ denote a fixed point of $\boldsymbol{\gamma} \circ \boldsymbol{\psi}$ and $\mathbf{u}^* = \boldsymbol{\alpha} \circ \boldsymbol{\beta}(\mathbf{v}^*)$. Since $\mathbf{v}^*$ belongs to $\mathcal{E}$, we have $\mathbf{v}^* = \boldsymbol{\gamma}(\mathbf{v}^*)$ so $\mathbf{u}^* = \boldsymbol{\psi}(\mathbf{v}^*)$ and $\boldsymbol{\psi}(\mathbf{u}^*) = \boldsymbol{\psi} \circ \boldsymbol{\psi}(\mathbf{v}^*)$. Remarking that

$$\boldsymbol{\psi} \circ \boldsymbol{\psi}(\mathbf{v}^*) = \boldsymbol{\psi} \circ \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v}^*) = \boldsymbol{\psi}(\mathbf{v}^*)$$

we get $\boldsymbol{\psi}(\mathbf{u}^*) = \boldsymbol{\psi}(\mathbf{v}^*) = \mathbf{u}^*$. ■

Proof of Lemma 2. d is a distance iff, for all $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ in $\mathcal{F}$, (i) $d(\mathbf{u}, \mathbf{v}) \geq 0$; (ii) $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$; (iii) $\{d(\mathbf{u}, \mathbf{v}) = 0\} \Leftrightarrow \{\mathbf{u} = \mathbf{v}\}$; (iv) $d(\mathbf{u}, \mathbf{w}) \leq d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w})$. (i), (ii) and (iii) are straightforward. (iv) is due to

$$\max_i \left(\frac{u_i}{w_i}\right) = \max_i \left(\frac{u_i}{v_i}\frac{v_i}{w_i}\right) \leq \max_i \left(\frac{u_i}{v_i}\right) \max_i \left(\frac{v_i}{w_i}\right)$$

and conversely for the min. ■

Proof of Lemma 3. The second case is obvious since d is a distance. So we concentrate on the proof of the first one. The main idea is to prove that $\boldsymbol{\beta}$ can not increase the distance d and that $\boldsymbol{\alpha} \circ \boldsymbol{\beta}$ necessarily reduces it.

$\boldsymbol{\beta}$**:** For $\mathbf{v} \neq \mathbf{w}$ we define $c_1 = \min_i(w_i/v_i)$ and $c_2 = \max_i(w_i/v_i)$. Remark that $c_1 < 1 < c_2$, $d(\mathbf{v}, \mathbf{w}) = \ln(c_2/c_1)$ and

$$\text{For all } i : c_1 v_i \leq w_i \leq c_2 v_i. \tag{6}$$

Denote $v'_j = \beta_j(\mathbf{v})$ and $w'_j = \beta_j(\mathbf{w})$. Since all the b_{ij} are positive, (6) implies that $c_1 v'_j \leq w'_j \leq c_2 v'_j$ for all j, which means that $\boldsymbol{\beta}$ does not increase d.

$\boldsymbol{\alpha} \circ \boldsymbol{\beta}$**:** Denote now $v''_j = \alpha_j[\boldsymbol{\beta}(\mathbf{v})] = v'_j/(1 + v'_j)$ and $w''_j = \alpha_j[\boldsymbol{\beta}(\mathbf{w})] = w'_j/(1 + w'_j)$. Remarking that the transformation $t \rightarrow t/(1 + t)$ is increasing, we derive that if $w'_i \geq v'_i$ then $w''_i \geq v''_i > c_1 v''_i$, and if $w'_i < v'_i$ then

$$w''_i = \frac{w'_i}{1 + w'_i} > \frac{w'_i}{1 + v'_i} \geq c_1 \frac{v'_i}{1 + v'_i} = c_1 v''_i.$$

So, in both cases, we have $w''_i > c_1 v''_i$ for every i.
Conversely, if $w'_i \leq v'_i$ then $w''_i \leq v''_i < c_2 v''_i$, and if $w'_i > v'_i$ then

$$w''_i = \frac{w'_i}{1 + w'_i} < \frac{w'_i}{1 + v'_i} \leq c_2 \frac{v'_i}{1 + v'_i} = c_2 v''_i.$$

So, in both cases, we have $w''_i < c_2 v''_i$ for every i.

This shows that, for every i, $c_1 v''_i < w''_i < c_2 v''_i$. So, denoting $c''_1 = \min_i(w''_i/v''_i)$ and $c''_2 = \max_i(w''_i/v''_i)$ we have $c_1 < c''_1$ and $c_2 > c''_2$, which implies that

$$d(\boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{w}), \boldsymbol{\gamma} \circ \boldsymbol{\psi}(\mathbf{v})) = \ln(c''_2/c''_1) < \ln(c_2/c_1) = d(\mathbf{w}, \mathbf{v}).$$

We conclude that $\boldsymbol{\alpha} \circ \boldsymbol{\beta}$ strictly reduces d. ■