

kerfdr: A semi-parametric kernel-based approach to local False Discovery Rate estimations

Mickaël Guedj^{*2,3,4}, Alain Céliisse^{1,4}, Gregory Nuel^{2,4}, Stéphane Robin^{1,4}

¹Statistics and Genome group, UMR AgroParisTech / INRA 518, Paris, FRANCE

²Statistics and Genome laboratory, UMR CNRS 8071 / INRA 152 / University of Evry, Evry, FRANCE

³Merck-Serono, Paris, FRANCE

⁴Statistics for Systems Biology Group, Paris, FRANCE

Email: Mickael Guedj* - mickael.guedj@gmail.com;

*Corresponding author

Abstract

Background: The use in Biology of current high-throughput genetic, genomic and post-genomic data leads to the simultaneous evaluation of a huge number of statistical hypothesis and at the same time, to the multiple-testing problem. As an alternative to the too conservative Family-Wise Error-Rate (FWER), the False Discovery Rate (FDR) has appeared for the last ten years as the more appropriate criterion to handle the problem. One drawback is that the FDR is associated to a given rejection region for the statistic considered, without distinguishing those that are close to the boundary and those that are not. As a result, the local FDR has been recently proposed to quantify the specific probability, given the p -value, for being a true-null hypothesis.

Results: In this context we present a semi-parametric approach based on kernel estimators which is applied to different high-throughput biological data such as genes expression and genome-wide association studies. Our estimation of the local FDR (ℓFDR) relies on the semi-parametric mixture model proposed in Robin et al (2007). We have at our disposal n hypotheses $\{H_i\}_{i=1,\dots,n}$ we want to test. Suppose that an unknown proportion π_0 of them are true nulls. For any hypothesis, we define a random variable H_i that equals 0 if it is under \mathbf{H}_0 (true null hypothesis), and equals 1 under \mathbf{H}_1 (false null). For each H_i , we compute a score denoted by X_i (a p -value for example). We assume that these scores are independent and identically distributed, with mixture distribution

$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x), \tag{1}$$

where $\pi_1 = 1 - \pi_0$ states for the proportion of false null hypotheses, f_0 denotes the probability density function (pdf) of scores under \mathbf{H}_0 and f_1 is the pdf of scores under \mathbf{H}_1 . Note that f_0 is completely specified. For instance if X_i is the p -value of a Student statistic, f_0 is the uniform distribution on $[0, 1]$. If any transformation (probit or log) is applied, f_0 remains completely known. On the contrary, f_1 needs systematically to be estimated so as to π_0 .

In our framework, ℓFDR defined the probability that $H_i = 0$ given the observed value x_i of the score X_i :

$$\ell FDR(x_i) \stackrel{def}{=} \tau_i = \Pr[H_i = 0 \mid X_i = x_i] = \frac{\pi_0 f_0(x_i)}{f(x_i)}.$$

This quantity may be interpreted as a measurement of how likely the hypothesis at hand could be falsely rejected.

Since f_1 is unknown, we use the following (nonparametric) kernel estimator for a given bandwidth $h > 0$

$$\hat{f}_1(x) = \left[\sum_{i=1}^n \frac{H_i}{h} k\left(\frac{x - X_i}{h}\right) \right] / \left(\sum_{j=1}^n H_j \right), \quad (2)$$

in which we replace the unknown H_i 's by their conditional expectation $\mathbb{E}[H_i \mid X_i] = \Pr[H_i = 1 \mid X_i] = 1 - \tau_i$.

These expectations are themselves thanks to

$$\hat{\tau}_i = \hat{\pi}_0 f_0(x_i) / \hat{f}(x_i), \quad (3)$$

where $\hat{\pi}_0$ is a given estimator of the unknown proportion and $\hat{f}(x) = \hat{\pi}_0 f_0(x) + (1 - \hat{\pi}_0) \hat{f}_1(x)$. Thus, we obtain

$$\hat{f}_1(x) = \left[\sum_{i=1}^n \frac{1 - \hat{\tau}_i}{h} k\left(\frac{x - X_i}{h}\right) \right] / \left(n - \sum_{j=1}^n \hat{\tau}_j \right). \quad (4)$$

As $\hat{\tau}_i$'s and \hat{f}_1 depend on each other, we alternate the computation of (3) and (4) until convergence, which is proved in Robin et al (2007).

Conclusions: The proposed method has the practical advantages, over existing approaches, to consider complex heterogeneities in the alternative hypothesis, to take into account prior information (from an expert judgment or previous studies) by allowing a semi-supervised mode and to deal with truncated distributions such as those obtained in Monte-Carlo simulations. Moreover, the method has been implemented and is now available through the R package `kerfdr`: <http://stat.genopole.cnrs.fr/software/kerfdr>.