`kerfdr`: A semi-parametric kernel-based approach to local FDR estimations

Grégory Nuel[1], **Mickaël Guedj**[1], **Alain Célisse**[2] **and Stéphane Robin**[2]

[1]Laboratoire Statistique & Génome (Evry, France) [2]INA-PG (Paris, France)

contact: `nuel@genopole.cnrs.fr`

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 quantifies the specific probability, given the p-value, for being under the null hypothesis. In this context we present a semi-parametric approach based on kernel estimators which we apply to different high-throughput biological data concerning patterns in DNA sequences, genes expression and genome-wide association studies. The proposed method has the practical advantages, over existing approaches, to consider both complex heterogeneities in the alternative hypothesis, prior information (from an expert judgment or previous studies) by allowing a semi-supervised mode and truncated distributions such as those obtained by using Monte-Carlo simulations. Moreover, the method has been implemented and available through the R package `kerfdr`.

Local FDR

Multiple-testing

Multiple-testing problems occur in many bioinformatic, genetic or bio-medical studies. The generic situation is the following: considering a large set of biological objects (genes, SNPs, DNA patterns, etc.), we want to test a null hypothesis H for each object. Typically, H = 'the expression level of the gene is not affected by the treatment' or 'the pattern as frequent as expected in the observed DNA sequence'. The control of the number of false-positives, i.e. hypotheses declared to be false whereas they are actually true is the crucial issue of multiple-testing. Many strategies have been proposed to control the Family-Wise Error-Rate (FWER) or the False Discovery Rate (FDR) see [1] for a review.

Mixture model

More recently, a strong attention has been paid to the estimation of the local version of FDR, called 'local FDR' [2] and denoted hereafter ℓFDR. The idea is to provide an estimation of the probability for a given hypothesis to be true or false. A very convenient way to define this notion consists in assuming that the testing scores are independant and identically distributed according to the following mixture distribution:

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

where π_0 (resp. π_1) is the proportion of true null (resp. true alternative) hypothesis $\mathbf{H}_0$ (resp. $\mathbf{H}_1$) in the sample, and where f_0 denotes the probability distribution function (pdf) of scores under the corresponding hypothesis.

Local FDR

The local FDR of the i-th observed value x_i of the score X_i is denoted by $\ell FDR(x_i)$ and defined as the probability that it is drawn from $\mathbf{H}_0$:

$$\ell FDR(x_i) \stackrel{def}{=} \tau_i = \Pr[H_0 \mid X_i = x_i] = \frac{\pi_0 f_0(x_i)}{\pi_0 f_0(x_i) + \pi_1 f_1(x_i)} = \frac{\pi_0 f_0(x_i)}{f(x_i)}\,.$$

This quantity may be interpreted as a measurement of how likely the alternative hypothesis $\mathbf{H}_1$ at hand could be falsely rejected.

Kernel-based estimation

An iterative approach

Since f_0 is perfectly known we only need to estimate both π_0 and f_1. This can be easily done in a parametric framework (gaussian mixture for example) but the critical choice of the parametric familly is often a difficult problem to address. As an alternative, we propose here use a very general kernel based estimator for f_1 thus able to capture any specific shape of the observed data. We hence get an iterative estimator ([3]):

$$\widehat{f}_1(x) = \left[\sum_{i=1}^{n} \frac{1-\widehat{\tau}_i}{h}\, k\left(\frac{x-x_i}{h}\right)\right] \Big/ \left(n - \sum_{j=1}^{n} \widehat{\tau}_j\right) \quad \text{and} \quad \widehat{\tau}_i = \frac{\widehat{\pi}_0 f_0(x_i)}{\widehat{\pi}_0 f_0(x_i) + (1-\widehat{\pi}_0)\widehat{f}_1(x_i)}$$

where k is a kernel function, h the bandwidth parameter and where $\widehat{\pi}_0$ have been previously estimated.

A R package: `kerfdr`

This method have been implemented in a `R` package called `kerfdr` which has the following features:

- a straightforward usage: `res=kerfdr(pv)` where `pv` is a sample of p-values returns the estimates of π_0 and ℓFDR in `res$pi0` and `res$localfdr`;
- quasi instant running time thanks to an efficient implementation using convolution through fast Fourier transforms;
- take into account censored input (especially useful when p-values are computed through Monte-Carlo estimations);
- semi-supervised mode allowing to take into account *a priori* knowledge on the studied sample;
- a powerful list of customizable options for advanced users (choice of π_0 or bandwidth estimator, of the kernel function, etc.).

Application to DNA patterns

We consider the complete genome of the pathogen bacteria *Mycoplasma genitallium* (575 kb) on which we estimate an order $m = 3$ homogeneous Markov model. For each of the $4^6 = 4,096$ oligomers of length 6, we compute the exact expectation ($\mathbb{E}[N]$) and standard deviation ($\sqrt{\mathbb{V}[N]}$) of its frequency N from which we derived the z-score:

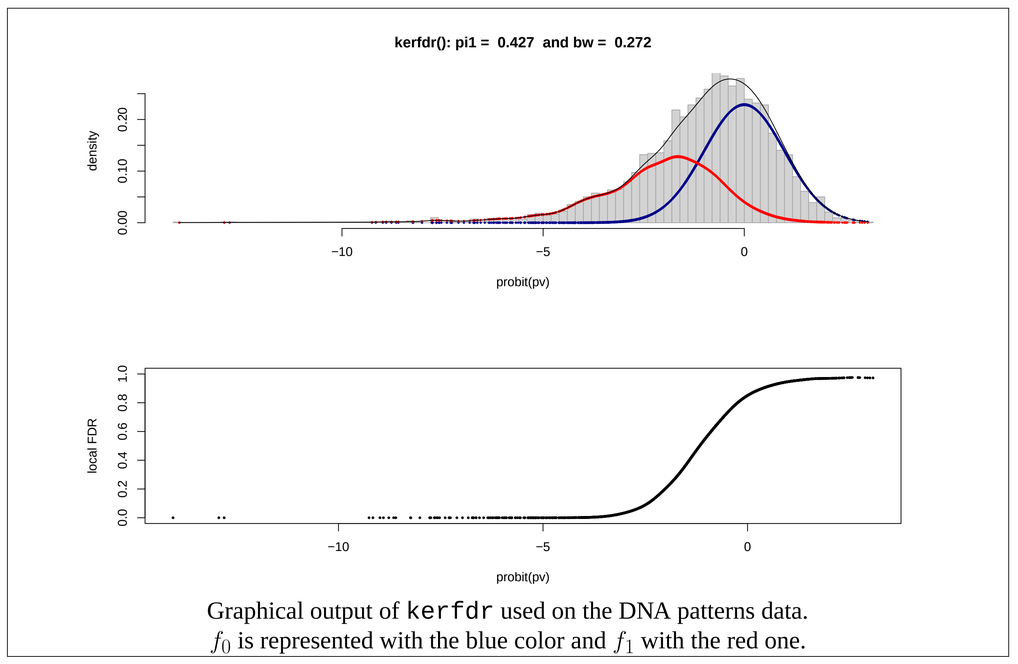
$$Z = \frac{N^{\text{obs}} - \mathbb{E}[N]}{\sqrt{\mathbb{V}[N]}} \underset{\mathbf{H}_0}{\sim} \mathcal{N}(0,1)$$

where N^{obs} is the observed frequency of the oligomer in the genome.

Thanks to a simple TCL argument, we got that the distribution of Z is approximately a standard Gaussian under the null hypothesis. It it hence possible to use this approximation by computing the two-sided p-value associated to each observation:

$$p\text{-value} = \mathbb{P}\big(\mathcal{N}(0,1) < -|Z|\big) + \mathbb{P}\big(\mathcal{N}(0,1) > +|Z|\big)$$

Our method is then applied on the resulting sample of p-values and the graphical output of `kerfdr` is available on the figure below.



Graphical output of `kerfdr` used on the DNA patterns data.
f_0 is represented with the blue color and f_1 with the red one.

Conclusions

`kerfdr` is a simple but powerful tool to deal with multi-testing problems. Thanks to its semi-parametric kernel based approch, the method does not assume any contrained alternative distribution and is hence far more flexible than purely parametric approaches. `kerfdr` also offers interesting extensions such as the possibility to use prior information in the estimation procedure (semi-supervised) and the ability to handle truncated distributions such as those generated by Monte-Carlo estimations of p-values for instance.

The corresponding `R` package will be soon made available in the CRAN and is already freely available at the following location:

`http://stat.genopole.cnrs.fr/softwares/kerfdr`

References

[1] Duboit S, Shaffer JP, Boldrick JC: **Multiple hypothesis testing in microarray experiments.** *Statistical Science* 2003, **18**:71-103.

[2] Efron B, Tibshirani R, Storey JD, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J. Amer. Statist. Assoc.* 2001, **96**:1151-1160.

[3] Robin S, Bar-Hen A, Daudin JJ, Pierre L: **A semi-parametric approach for mixture models: application to local false discovery rate estimation.** *Comput. Statist. and Data Analysis* 2007, **51**:5483-5493.